DOCUMENT RESUME

ED 124 541 95 SP 010 157

AUTHOR Goulet, Larry R.; And Others
TITLE Investigation of Methodological Problems in Educational Research: Longitudinal Methodology. Final Report.
INSTITUTION Illinois Univ., Urbana.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
BUREAU NO BR-4-1114
PUB DATE Sep 75
CONTRACT NIE-C-74-0124
NOTE 261p.

EDRS PRICE MF-$0.83 HC-$14.05 Plus Postage.
DESCRIPTORS *Behavior Change; *Educational Research; *Equated Scores; *Longitudinal Studies; *Measurement Techniques; Mental Development; Research Design; *Research Methodology; Standardized Tests; Testing; Test Interpretation; Time Factors (Learning); True Scores
IDENTIFIERS *Anchor Test Study

ABSTRACT

The problems and issues involved in the conduct of educational-developmental research are examined within the perspective of longitudinal research methodology. Chapters 2 and 3 examine contemporary research designs and procedures implemented for the selection of subjects and testing of behavior over time. Particular attention is given to the sequential research paradigms developed by Schaie for the purpose of simultaneous assessment of age-, cohort-, and time-related behavior change. Some of the common problems in measuring change and models for analyzing longitudinal data are considered in Chapters 4 through 7. Particular emphasis is given to interpretive problems resulting from the properties of scales widely used on standardized achievement tests, to the limitations of current techniques of vertical equating and consideration of alternative equating methods, and to the evaluation of the constancy of a construct over time. Chapter 8 presents an exposition of time-series analysis along with a new procedure for parameter estimation especially adapted to data from longitudinal studies. In Chapter 9 the problems of measurement of true change are reconsidered, and it is stated that lower and upper bounds for estimated true change are derived under more relaxed conditons than in classical test theory. The appendix reviews the report of the Anchor Test Study. (Author/MM)

FINAL REPORT

PROJECT NO. 4-1114
CONTRACT NO. NIE-C-74-0124

INVESTIGATION OF METHODOLOGICAL PROBLEMS IN EDUCATIONAL RESEARCH: LONGITUDINAL METHODOLOGY

LARRY R. GOULET
ROBERT L. LINN
MAURICE M. TATSUOKA

UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN, ILLINOIS

SEPTEMBER, 1975

The research reported herein was performed pursuant to a contract with the National Institute of Education, Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgement in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official National Institute of Education position or Policy.

U.S. DEPARTMENT OF HEALTH,
EDUCATION, AND WELFARE

NATIONAL INSTITUTE OF EDUCATION

FINAL REPORT

PROJECT NO. 4-1114
CONTRACT NO. NIE-C-74-0124

INVESTIGATION OF METHODOLOGICAL PROBLEMS IN EDUCATIONAL RESEARCH-LONGITUDINAL METHODOLOGY

LARRY R. GOULET
ROBERT L. LINN
MAURICE M. TATSUOKA

UNIVERSITY OF ILLINOIS AT
URBANA-CHAMPAIGN, ILLINOIS

SEPTEMBER, 1975

U.S. DEPARTMENT OF HEALTH,
EDUCATION, AND WELFARE

NATIONAL INSTITUTE OF EDUCATION

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The research team that worked on this project consisted of the three principal investigators, Larry R. Goulet, Robert L. Linn, and Maurice M. Tatsuoka, a research associate, Kikumi K. Tatsuoka, who joined the project in January 1975, and three research assistants, Craig Barclay, Jeffrey A. Slinde and Michael Townsend. For most of the duration of the project, Patsy M. Rowland served as the project secretary.

Each of the principal investigators took responsibility for particular problem areas that were addressed in this project. The areas of primary responsibility are reflected in the chapters of this report. The writing responsibility for chapters 2 thru 9 and for Appendix A is as follows:

Chapter 2: Larry R. Goulet, Craig Barclay, and Michael Townsend
Chapter 3: Larry R. Goulet, Craig Barclay, and Michael Townsend
Chapter 4: Robert L. Linn and Jeffrey A. Slinde
Chapter 5: Robert L. Linn and Jeffrey A. Slinde
Chapter 6: Robert L. Linn
Chapter 7: Robert L. Linn
Chapter 8: Maurice M. Tatsuoka
Chapter 9: Kikumi K. Tatsuoka and Maurice M. Tatsuoka
Appendix A: Robert L. Linn

Particular thanks go to Patsy M. Rowland for her care and speed in typing several drafts, as well as part of the final report. Chapters 8 and 9 were typed by Mrs. Joyce Sterner of Technitypists, Inc., Urbana, Illinois.

CHAPTER 1

INTRODUCTION

The basic premise upon which this report rests is that the development and advancement of theory in education, the generation of data and theory directly relevant to school programs and individual classrooms, and the opportunity to examine complex educational questions await the development of an appropriate methodology. Such a premise is similar to that made by George Mandler (1967) in discussing contemporary approaches to the experimental study of learning processes. He suggested, for example, that contemporary research on human learning emphasizes an "active," rather than a "passive" organism, and a shift to the study of "complex" processes -- without the necessity of conducting "complex experiments." The latter coup was attributed to the development of and advances in our knowledge concerning research methods.

Similar types of comments have been made by Fiske (1973) in discussing the need for process-type research in the personality area. He suggests, "the central but only vaguely recognized need is for intensive work on the basic strategy of psychological research, especially in the personality domain," and further asks, "can we study the important psychological processes in the laboratory or testing room? How can we be sure of the occurrence of the postulated process? Or do we define each specific process simply as that which we presume to occur between a particular stimulus and a designated type of response." Fiske also suggests that laboratory research, in addition to facing problems regarding the replicability of process-type phenomena, faces an almost insurmountable problem -- that of determining the degree to which the findings are generalizable to behavior in general.

Wohlwill (1973) has also addressed such questions from the perspective of developmental psychology. In addressing the question whether developmental research belongs in an "experimental" or "differential" camp, he suggests, "it turns out that the study of developmental change does not readily fit either of the two models, at least in their simplest form. On the one hand, the study of age changes in behavior differs, in certain important respects, from comparative, differential investigations involving other interpersonal characteristics, e.g., the study of sex differences. On the other hand, even when development is subjected to direct experimental attack by manipulating the conditions of experience in a controlled manner, the situation still deviates in some critical ways from that which confronts the experimentalist dealing with nondevelopmental problems. Thus, the concern with development gives rise to very particular requirements and considerations as regards experimental methodology, research design, and scientific inference. To put it succinctly:

The canons of the scientific method, as they have been worked out for the field of psychology at large, require modification when applied to developmental problems." (16-17)

The comments of Mandler (1967), Fiske (1973), and Wohlwill (1973) are equally appropriate to educational research, not only because of the partial overlap of content across these disciplines, or the common call for the development of new methods, but also because each has called for the study of the respective phenomena in the environmental contexts in which they occur and because each calls for the further development of research methods which provide for direct, unconfounded, and generalizable estimates of these processes as they change with time.

LONGITUDINAL AND CROSS-SECTIONAL METHODOLOGY

Some History of and the Interdisciplinary Character of Longitudinal Research

As Sontag (1971) has noted, longitudinal methodology is by no means under the exclusive purview of developmental psychology. Its roots are found in a variety of disciplines including demography and multiple social sciences, life sciences, and physical sciences. Yet, he suggests that the term longitudinal research evokes free associations of a "womb-to-tomb" research plan, inadequate research design, inexact measurement, and an inadequate and inordinately expensive research product. Yet, and somewhat paradoxical, the longitudinal method and the superiority of longitudinal data over cross-sectional data, remains essentially unquestioned in educational and developmental research; e.g., Hilton & Patrick (1971). Similarly, cross-sectional methodology is seen primarily as a convenient but approximate substitute for longitudinal measurement.

The qualms of scientists regarding the use of longitudinal designs can be traced to a number of relevant problems. For example, the use of a longitudinal design usually requires that the experimenter "age" with his subjects, the fact that the experimenter cannot control the subjects' experiences between the several times of testing, subject attrition, and perhaps more important, the fact that the longitudinal method commits the experimenter to a specific design and the use of specific measurement instruments over the duration of the study.

Such difficulties have been noted as early as 1741 by Sussmilch who also, by the way, commented on the problems of generalizability of using what we now call cross-sectional methods.

Quetelet (1835) and Galton (1883) were advocates of the cross-sectional method, yet it was not until the 1920's that the terms longitudinal (Blatz & Blott, 1927) and cross-sectional (Gesell, 1925) were used to designate the different methods, and it was Anderson (1931), in his classic contribution to developmental methodology, who affirmed their use as technical terms.

Considering the importance and use of longitudinal and cross-sectional methodology in educational and developmental research, it is unfortunate, and surprising that comprehensive and satisfactory discussions of the problem are unavailable in the educational literature. As an example, it is in demography where significant advances have been made (e.g., Whelpton, 1954). The lack of consideration of these advances in other disciplines is particularly unfortunate since, as one case in point, large-scale educational research related to student development has borrowed conventional designs only from developmental psychology rather than likely more appropriate adaptations of these designs used by the demographers.

"Experimental" and "Descriptive" Designs and Variables

Parallel types of criticisms have been directed to studies utilizing longitudinal and cross-sectional sampling designs. The primary criticism relates to the difficulty in assigning causality or the directionality of relationships in such studies (Campbell & Stanley, 1963; Russell, 1957; Spiker, 1966) and the inability to subscribe fully to the principles of experimental design when these procedures are used. As an example, chronological age is a biotic variable not amenable to random assignment, replication, etc. Yet, invoking the principle that only properly randomized experiments can lead to useful estimates of causal treatment effects, is a potential trap for educational researchers. As examples, it may lead educational researchers to reject one of the primary (if not the primary) problem in the field -- i.e., the estimation of the influences of educational (e.g., classroom) experiences on performance; it can lead to the design of educational research blindly following the principles of experimental design at the expense of the crucial focus -- the critical analysis of educational environments and the attendant individual-environment interactions. It also encourages "laboratory" investigations rather than studies which take place in the less-controlled educational context. And, it encourages investigations where data are collected at one time of measurement rather than longer-term studies and possible sacrifices in external validity for gain in internal validity.

In addition, the costs, in terms of time and money are indeed prohibitive when "experiments" are conducted (Rubin, 1972). This is true since it is impossible to perform equivalent experiments to test all treatments on even a single educational question (e.g., examining 100 reading programs). And, the above argument has not included the argument that the exclusive use of experimental variables precludes the study of certain educational questions or that rnadom assignment cannot be ethically used as a proeedure in certain types of studies.

Several of the questions and issues discussed above relate to questions of research design and methodology and are addressed in Chapters 2 and 3.

Specific Methodological Problems in Longitudinal Research

Longitudinal studies confront numerous difficulties, only a fraction of which were addressed within the confines of this project. A variety of issues involved in the measurement of change are considered in Chapter 4. Of particular concern in Chapter 4 are difficulties caused by characteristics of scales commonly used for standardized achievement tests.

Studies, whether longitudinal or cross-sectional, which focus on student achievement over a period of several years typically require different measures of achievement at different grades or ages. In order to make comparisons of achievement over time such tests must be put on a common scale, i.e. they must be vertically equated. In Chapter 5 the adequacy of the vertical equating of some existing standardized achievement tests is investigated and a study exploring the potential utility of the Rasch model for the vertical equating problem is reported.

Several attempts at using analytical techniques developed by Joreskog for the analysis of covariance structures are discussed in Chapters 6 and 7. In Chapter 6 the focus is on the fit of several sets of data to a Simplex model and in Chapter 7 the focus is on the use of these techniques to evaluate the constancy of constructs over time.

Time-Series Analysis in Longitudinal Research

From its very name, time-series analysis seems to be a technique especially suited to longitudinal research. A casual study of its methodology reveals, however, that--as traditionally conducted--it is applicable more to sequential cross sectional research. In Chapter 8 we first present an elementary exposition of time-series analysis, then indicate the difficulties in applying it to data from longitudinal studies as ordinarily conceived, and finally propose a new method for estimation of parameters in time-series models that is especially adapted to longitudinal data.

In brief, the difficulties with the traditional procedures for parameter estimation in time-series analysis are that (a) they require a large number ($\geq$ 50) of time-point observations, and (b) they ignore the correlatedness of individual data across time. A procedure which avoids these difficulties is proposed and successfully tested by means of two numerical examples, one based on real data and the other using simulated data.

Measurement of Change

The time-honored problem of measurement of time change is revisited in Chapter 9. Difficulties with the traditional assumption of "universally uncorrelated errors" are discussed in this context,

and a relaxed assumption of "homogeneity of error covariances" is proposed. Under the latter assumption, lower and upper bounds for estimated time change are derived, utilizing the mathematics of operator analysis.

An example based on real data is presented, and it is shown that the uncorrelated-errors assumption leads to an absurd result (a multiple-R greater than unity), while the relaxed condition yields reasonable and useful bounds.

REFERENCES

Anderson, J. E. The methods of child psychology. In C. Murchison Ed.), A handbook of child psychology. Worcester: Clark University Press, 1931.

Blatz, W. E., & Blott, E. A. Studies in mental hygiene: I. Behavior of public school children --description of method. Pedagogical Seminary, 1927, 34, 552-582.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 171-246.

Fiske, D. W. Research on Psychological processes with particular reference to personality. In S. B. Sells (Ed.), Needed research on psychological processes. Washington, D. C., U. S. Office of Education, 1973.

Galton, F. Inquiries into human faculty and its development. London: MacMillan, 1883.

Gesell, A. L. The mental growth of the pre-school child: A psychological outline of normal development from birth to the sixth year, including a system of developmental diagnosis. New York: Macmillan, 1925.

Hilton, T. L., & Patrick, C. Cross-sectional versus longitudinal data: An empirical comparison of mean differences in academic growth. Journal of Educational Measurement, 1970, 7, 15-24.

Mandler, G. Verbal learning. In G. Mandler, P. Mussen, N. Kogan, and M. A. Wallach, (Eds.) New Directions in Psychology III. New York: Holt, Rinehart and Winston, 1967.

Quetelet, A. L. Sur l'homme el le developpement de ses facultes. Paris: Bachelier, 1835.

Rubin, D. Estimating causal effects of treatments in experimental and observational studies. Princeton, N. J.: Educational Testing Service, Research Bulletin 72-39, 1972.

Russell, W. A. An experimental psychology of development: Pipe dream or possibility. In D. B. Harris (Ed.), The Concept of Development. Minneapolis: University of Minnesota Press, 1957, 162-174.

Sontag, L. W. The history of longitudinal research: Implications for the future. Child Development, 1967, 42, 987-1002.

Spiker, C. C. The concept of development: Relevant and irrelevant issues. In Stevenson, H. W. The concept of development. Monographs of the Society for Research in Child Development, 1966, 31, No. 2 (Serial No. 107), 40-54.

Whelpton, P. K. Cohort fertility (native white women in the United States). Princeton: University Press, 1954.

Wohlwill, J. The study of behavioral development. New York: Academic Press, 1973.

CHAPTER 2

THE STUDY OF BEHAVIOR CHANGE OVER TIME

OVERVIEW

The study of time-related behavior change comes in varied forms. To the developmental psychologist, such a research focus most typically implies the study of behavioral development. For the sociologist, such a purpose more likely would imply the study of social or sociocultural change. The educational researcher is concerned with each of these in a very direct way. We are concerned with how the population of school children changes across time, e.g., years or decades, and with the performance changes of specific groups of children as they pass through successive school grades. The first two purposes notwithstanding, the educational researcher is often confronted with a third and more specific question, i.e., the assessment of the influences of schooling or educational intervention. The differences, similarities and interrelationships among these various research questions are discussed in detail in various sections of Chapters 2 and 3. It is our intention to examine various research designs and theoretical models which fit such questions. Several theoretical assumptions which underly these questions are also examined most specifically as they apply to longitudinal methodology. These questions and designs are discussed in this chapter in the context of conventional procedures and sampling methods. Modifications and extensions of such designs proposed by Bell (1953), Schaie (1965) and Baltes (1968) are presented and discussed. In Chapter 3, general sampling procedures are presented which can be adapted to the theoretical model and assumptions adopted by the researcher.

Many of the inferences of this paper rest on the assumption that educational research, like developmental research, can be described by problems which take the form:

$$B = f(T)$$

where "B" refers to the behavior or behavior changes to the studied, and "T" refers to the time period over which the assessments are made (Baltes and Goulet, 1971). As will be shown, most designs used in educational research can be described by the above paradigm even though they represent only the simplest cast of a more general model for research concerned with changes in behavior associated with time. These research designs are discussed and their limitations in the context of educational research are noted in the next section.

SIMPLE DESIGNS FOR EDUCATIONAL RESEARCH

Schaie (1965) has noted that the paradigm B = f(T) described above spawns three alternate research designs, generally known as the cross-sectional method, the longitudinal method, and the time-lag method. These three designs differ in terms of the procedures used to draw the samples of interest and the time period over which measurements are taken. With the cross-sectional design, for example, samples of different ages are tested at the same point in time. As will be shown, such a design has limited usefulness in educational research. The longitudinal method requires the testing of samples with the same birthdate (or alternately samples who are in the same school grade) at different points in time. Such a design is perhaps the most popular of the three in the context of educational research since the children can be followed over periods of time when they are enrolled in school.

It is important at this point to mention that the longitudinal design is amenable to both between-S and within-S (i.e., repeated measurement) testing procedures. As mentioned above, the basic requirements of the longitudinal design are met if Ss with the same birthdate are tested at two or more points in time. This may be accomplished through the repeated testing of the same sample of Ss; i.e., a within-S longitudinal design. With a between-S longitudinal design, samples of Ss can be randomly drawn from a population born within the same period, with each sample being assigned to testing at one of the times of measurement represented in the investigation.

The time-lag design, the least used in educational research, yet perhaps the most powerful of the three designs for educational purposes, requires the testing of samples with different birthdates at the same chronological age. This, of course, requires testing the samples in the order in which they are born.

These three designs are represented in Figure 1, with the cross-sectional (Xs) design conforming to the vertical (cross-row) comparisons, the longitudinal (Lo) design conforming to the horizontal (cross-column) comparisons, and the time-lag (Tl) design conforming to the diagonal comparisons. As Figure 1 also illustrates, a particular sample of Ss is fully described by three components, date of birth (cohort), age, (A) and time of testing (Schaie, 1965). Note, however, that the sampling model described in Figure 1 makes no reference to the level of educational attainment (e.g., school-grade) of the respective samples of subjects defined by the model. It is apparent that any prototypic design for educational research must provide for the estimation of such a parameter and this is discussed in later sections of the chapter. However, at this point it is most relevant to contrast the three alternate designs as they incorporate this parameter into one of the three already described.

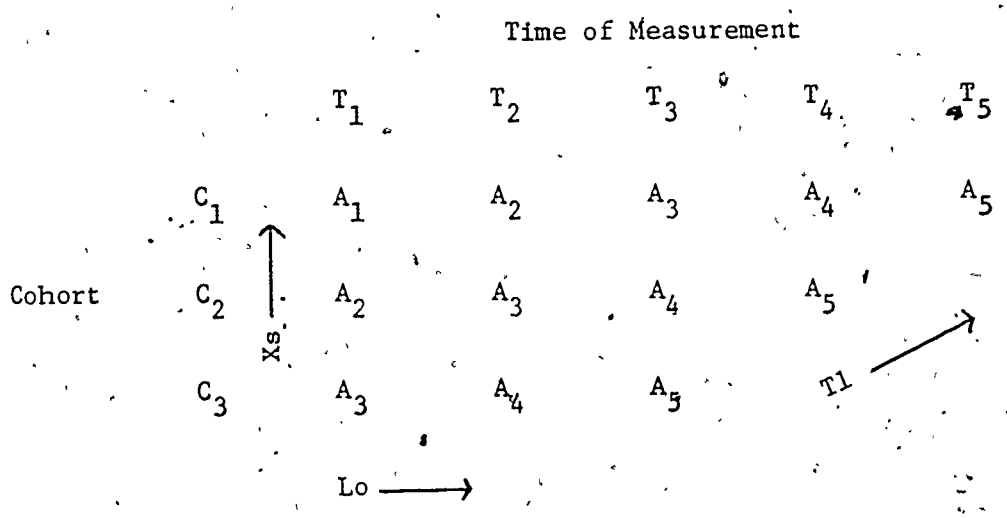


Figure 1

Simple Designs for Educational Research

Educational Attainment and the Cross-Sectional Design

Studies concerned with educational phenomena and utilizing the cross-sectional sampling procedures implicitly or explicitly incorporate educational experiences as part of the age component. Examples are studies where the samples of Ss tested differ in CA by a minimum of one year or a minimum of one school grade. As is apparent, such a procedure yields results which confound amount of schooling and other components of CA-related behavior change and, thus, the effects of educational experiences can be estimated only in conjunction with these other factors. Furthermore, the cross-sectional method requires the added assumption that the effects of schooling for children in comparable grades are the same irrespective of the year in which the children are enrolled. Thus assumption is similar to that made in developmental research; i.e., that measures of performance utilizing cross-sectional sampling procedures will provide results identical to those involving longitudinal sampling procedures (Wohlwill, 1970).

Similarly, within-grade cross-sectional contrasts (where between-CA contrasts are made for Ss in the same grade) have little use in educational research since this design does not provide for variation in the educational experiences of the samples.

Educational Attainment and the Longitudinal Design

The longitudinal design suffers from the same limitations as the cross-sectional method, except that the limitation holds when both within- and between-grade contrasts are made. Again, amount of school experience and other CA-related influences on behavioral development are inextricably correlated. In fact, the case has been made (Goulet, Williams, & Hay, 1974) that, because of the confounding of CA-related and school-related influences on development, the longitudinal method will normally provide estimates of behavior change which exceed those involving the cross-sectional method when within-grade contrasts are made.

Educational Attainment and the Time-Lag Method

In contrast to the cross-sectional and longitudinal methods, the use of the time-lag methods, perhaps more properly, identifies school experience with the time-of-testing component in Figure 1. The use of this design in educational research, although somewhat limited by the age-graded nature of the schools, nevertheless permits both within-grade and between-grade contrasts to be made for samples of varying CAs. The design capitalizes on two simple facts; i.e., that children within a grade differ in CA, and that the CA of a sample of Ss increases over the period of a school year. Thus, in reference to Figure 1, if testing takes place in October and April within the same academic year, it is possible to contrast matched CA samples within a grade (e.g., at age A_2) or between matched-CA samples in adjacent school

grades (e.g., at age A_2). Such contrasts permit the estimation of the effects of school experiences independently of other CA-related factors. The major limitation of using the time-lag method is that the contrasts may only be made for Ss in adjacent grades or for within-grade contrasts. Nevertheless, many such contrasts can be made.

It is apparent that the use of a cross-sectional sampling strategy is inappropriate when the purpose of the researcher is in assessing education-related performance changes associated with time. The difficulty is further compounded when it is taken into consideration that cross-sectional differences in performance are as likely attributable to population (i.e., cohort) differences as to age differences, Bell (1953) and Kessen (1960) have each noted this possibility and have advocated the use of longitudinal sampling whenever population differences/changes are a possibility. However, longitudinal sampling, where Ss are repeatedly tested, suffer from potential contamination due to repeated observation, attrition, etc. Longitudinal measurement also "takes time" since the researcher must wait between successive testing periods. In addition, it is evident in Figure 1 that longitudinal changes in performance may be attributable to factors associated with age, time-of-testing, or both.

Bell's Convergence Method

Such difficulties in interpretation of the $B = f(T)$ functions have led to several suggested modifications of the above sampling procedures. The first of these was presented by Bell (1953) and called the Convergence Method. A prototype of the Convergence Method is presented in Figure 2.

Figure 2 describes four samples of children (cohorts 1962, 1964, 1966, 1968) each tested in three consecutive years (1974, 1975, 1976) and involves combining the longitudinal and cross-sectional sampling methods in such a way that "developmental changes for a long period may be estimated in a much shorter period (Bell, 1953, p. 147)." In other words, the age function from 6-14 in Figure 2 can be described by using three testing points (spanning a two-year period) for each of the four cohorts. The overlap in CA for the successive cohorts (e.g., cohorts 1968 and 1964 are each tested at the age of eight) is built into the design in such a fashion as to permit the possibility of assessing population differences. In other words, in the absence of performance differences across different cohorts matched on CA, Bell (1953) suggested that the longitudinal function estimated using the convergence method would overlap with the longitudinal function which would have been obtained if the 1962 cohort would have been tested at the age of six and yearly thereafter.

Bell's (1953) Convergence Method was suggested as an alternate sampling procedure (replacing longitudinal or cross-sectional methods) to reduce some of the difficulties associated with longitudinal sampling. Implicit in suggesting the method was the suggestion that

longitudinal sampling was clearly the method of choice when the purpose of the researcher is to describe developmental-age functions for a specific cohort or population of subjects.

Furthermore, Bell clearly anticipated recent refinements in longitudinal methodology by suggesting that combinations of longitudinal and cross-sectional sampling have merits which clearly exceed those using either sampling method alone. And, his suggestions have been tacitly accepted by Schaie (1965), Baltes (1968), Buss (1973), Goulet, Hay, & Barclay (1974), in recent papers which have had the primary purpose of identifying the components of time-related behavior change.

SEQUENTIAL METHODOLOGY

Schaie (1965) has criticized the available sampling methods and has suggested that longitudinal and cross-sectional methods are only special cases of a general model for research on behavior change over time. He argued that performance is a function of three factors, the age (CA) of the organism, the cohort (C), to which the organism belongs, and the time (T) at which measurement occurs, i.e., $R = f(A,C,T)$. A cohort, according to Schaie (1965) refers to the population of organisms born at the same point or interval in time. In short, Schaie (1965) suggested that differences associated with age which are obtained using longitudinal and cross-sectional sampling procedures would accurately reflect behavioral development (and provide identical estimates of age-related behavior change) only if there were no population (i.e., generation) or environmental (culture) changes over time. In the absence of evidence to the contrary, cross-sectional differences in performance must be assumed to reflect the combined influences of developmental (i.e., age) and population (i.e., cohort) changes associated with time. Similarly, longitudinal differences in performance reflect influences of age- and time-of-measurement-related factors.

In view of the potential confounding, Schaie proposed a model for the conduct of developmental research which provides the opportunity to examine the influences of each of these components on performance. The general model generates three different sequential research designs which permit CA, cohort, and time of measurement to be simultaneously varied, two at a time. The general model is summarized in Figure 3.

As Figure 3 indicates, samples of Ss representing five levels of age and nine cohorts are tested at five times of measurement. Between-row contrasts represent conventional cross-sectional (x-s) comparisons. Diagonal contrasts conform to a time-lag (Tl) design, and those between-column comparisons represent longitudinal (Lo) contrasts. As is apparent, cross-sectional comparisons confound age and cohort differences, longitudinal comparisons confound age and time of measurement differences, and time-lag comparisons confound cohort and time-of-measurement differences. In view of such confounding, Schaie (1965) suggested the use of sequential sampling designs which separate sources of variance associated with the three components. Thus, a cohort-

Cohort	Time of Testing 1974	1975	1976
1968	6[a]	7	8
1966	8	9	10
1964	10	11	12
1962	12	13	14

a. Cell entries refer to CA at the time-of-testing

Figure 2

Bell's Convergence Method

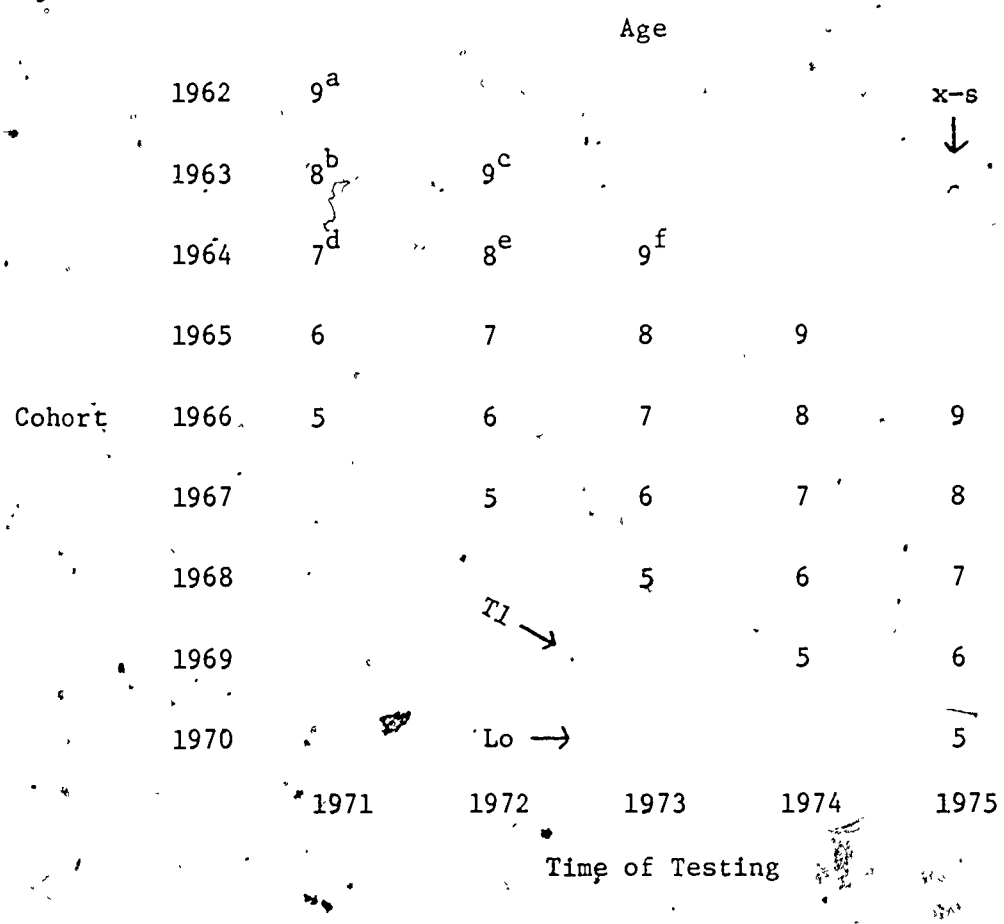


Figure 3

A Prototype of Schaie's General Developmental Model

sequential design, represented by samples b, c, e, and f in Figure 3 provides an estimate of age differences controlled for cohort differences and for cohort differences controlled for age. Similarly, a time sequential design represented by samples a, b, c, and e in Figure 3 provide for estimate of age differences with time of measurement controlled, and for time of measurement differences with age controlled. The cross-sequential design, represented by samples b, c, d and e in Figure 3, provide for estimates of cohort changes unconfounded by time and for time differences unconfounded by cohort differences. Schaie (1965) suggests further that a sampling plan conforming to the example provided in Figure 3 provides the opportunity to assess the independent effects of each of the three components with a minimum of six samples of Ss, e.g., samples a, b···g in Figure 3.

The primary ability of Schaie's model is that it provides methods for separating sources of developmental change. That is, unlike the cross-sectional method, the use of the cohort-sequential design provides the opportunity to examine age differences in the absence of confounding with the cohort variable. Similarly, the time-sequential design provides the possibility of identifying age-related effects without the confounding of time-of-measurement (as with the longitudinal method).

Nevertheless, the model as discussed to this point remains exclusively descriptive and no theoretical meaning can be ascribed to either age, cohort, or time-of-testing effects obtained when using the model. Schaie (1965) has, in fact, suggested that the three components are subject to theoretical interpretation that is, age differences estimated from the model may, according to Schaie be interpreted as the "net effect of maturational change," time differences as "net changes within the environment" and cohort effects as "net changes between generations" (1965, p. 96). Schaie suggests further that these effects may be estimated simultaneously, whenever data are available which conform to the general model; e.g., the six samples (a - g) in Figure 3.

The theoretical interpretations of age, time, and cohort effects proposed by Schaie have evoked considerable controversy (e.g., Baltes, 1968; Buss, 1973; Wohlwill, 1973). Baltes (1968), for example, has suggested that the three components, age (A), Time (T) and Cohort (C) do not exist independently of one another; i.e., that Schaie's model can be described adequately by two rather than three components. In other words, once two of the components are specified, the third is unequivocally fixed. This fact can be demonstrated by recourse to a simple example; i.e., that the cohort for any particular sample of Ss may be determined by subtraction of age (years) from the time of measurement; i.e.,

$$C = T - A \qquad 2.1$$

Similarly it can be shown that the following two relationships exist:

$$T = A + C \qquad 2.2$$

$$A = T - C \qquad 2.3$$

Baltes (1968) suggested that the existence of the mutual dependencies reduce the model to a bifactor rather than a trifactor model and that one of the components in Schaie's (1965) formula, $R = f(A, C, T)$, can be replaced by substitution. As an example, the substitution of $A + C$ in formula 2.2 Schaie's formula becomes $R = f(A, C, A + C)$. Further difficulties relating to the theoretical interpretation of $B = f(T)$ phenomena are discussed later in this chapter.

It is, however, important to consider the implications of Baltes' suggestions as they relate to research methodology and the adequacy of available sampling methods. First, in the absence of the possibility of functionally separating age from time-of-measurement effects, or cohort from time-of-measurement effects, the longitudinal and time-lag sampling methods immediately become (contrary to Schaie's suggestions) acceptable research designs for the study of $B = f(T)$ phenomena. These two designs are only limited in their generalizability; i.e., longitudinal data collected on a single cohort provide "true" estimates of age-related development for the cohort and time interval being studied. Similarly, performance differences estimated using the time-lag method provide true estimates of cohort-related change for the ages and time interval being studied. Only the cross-sectional sampling method is unacceptable since it confounds age and cohort effects.

The problem of generalizability is also reduced, according to Baltes (1968), if the longitudinal method is supplemented by: (1) obtaining longitudinal measurements for more than one cohort; i.e., by using the cohort-sequential design (or what Baltes calls longitudinal sequences); or (2) obtaining cross-sectional measurements across several times; i.e., using Schaie's time-sequential design (or what Baltes calls longitudinal sequences). Most important, both Schaie and Baltes recommend the use of sequential designs for the study of $B = f(T)$ phenomena and their use is most strongly recommended here whenever the intent of the researcher is to obtain acceptable (and generalizable) estimates of age or generation effects. It is apparent, however, that the estimates of $B = f(T)$ phenomena using sequential methods remain descriptive and subject to differing theoretical interpretations. The Baltes (1968) and Schaie (1965) controversy is a case in point.

SEQUENTIAL DESIGNS AND EDUCATIONAL RESEARCH

The above discussion has highlighted the difficulties in using conventional sampling methods in research oriented to the assessment of the influences of educational experiences. This discussion leads to three questions concerned with these problems.

1. Do children of the varying CA's enter a school grade with varying proficiency?

2. What are the non-CA-related influences of schooling?

3. What is the nature of the interaction between amount of schooling and CA in performance?

Unfortunately, none of the above designs previously discussed provide information concerning these questions. Nevertheless, it is possible to set up a sampling procedure which when used, permits these questions to be addressed directly. Figure 4 provides a prototype of such a sampling plan. In the figure, samples of Ss varying in CA (A_1, $A_2 \cdots A_8$); amount of schooling (S_1, $S_2, \cdots S_7$), and school grades are tested at different points in time during the period of a school year and permits cross-sectional contrasts (between-row comparisons) longitudinal contrasts (diagonal comparisons) and time-lag contrasts (between-column comparisons).

The cross-sectional contrasts (relevant to question 1 above) provide comparisons of performance for samples of children varying in CA but who have had the same amount of formal schooling. For the time-lag contrasts (relevant to question 2 above), the comparisons are for samples matched on CA who vary in amount of schooling. The longitudinal contrasts (where samples of Ss born during the same period are tested at different points in the school year) inextricably confound CA and amount of schooling. Fortunately, the cross-linking of appropriate samples (as exemplified in Figure 4) permits comparisons which provide information to be collected regarding each of the above three questions in the same analysis. For example, statistical contrasts involving samples a, b, c, and d in Figure 4 permit the behavior changes related to the first four months of schooling, CA, and their interaction to be estimated for children in first grade. An analysis involving samples d, e, f, and g from Figure 4 permits similar comparisons for the last four months of the school year. Finally, an analysis involving samples g, h, i, and j from Figure 1 permits educational growth during the latter part of first grade and the early part of second grade to be estimated. Each of the statistical analyses outlined above represent simple 2 x 2 factorial designs with CA and time of testing as the two factors. Furthermore, each analysis permits two independent assessments of the influences of schooling (one at each of two levels of CA) and two estimates of the relation between CA and performance (one at each of two times of testing in the school year). Additional discussion of the statistical analyses which follow from the use of the sampling plan in Figure 4 is presented later in this paper. However, it is important at this point

Chronological Age	Time of Testing: Sept. (Grade 1)	Jan. (Grade 1)	May (Grade 1)	Sept. (Grade 2)	Jan. (Grade 2)	May (Grade 2)
6-2	A_1 S_1					
6-6	A_2 S_1^a	A_2 S_2^b				
6-10	A_3 S_1^c	A_3 S_2^d	A_3 S_3^e			
7-2		A_4 S_2^f	A_4 S_3^g	A_4 S_4^h		
7-6			A_5 S_3^i	A_5 S_4^j	A_5 S_5	
7-10				A_6 S_4	A_6 S_5	A_6 S_6
8-2					A_7 S_5	A_7 S_6
8-6						A_8 S_7

Figure 4

A Sequential Sampling Procedure for Educational Research

to note that the samples of Ss represented in the present model are independent groups. Thus all comparisons conforming to cross-sectional, time-lag or longitudinal designs are based on between-S (as opposed to within-S) comparisons. Such contrasts may be made across the entire period of formal schooling and, interestingly, data conforming to the sampling plan in Figure 4 and spanning several school grades may be collected over the period of a single school year (e.g., 1975) or multiple school years, e.g., 1975, 1976···.

Descriptive Uses of the Sampling Procedure in Figure 4

The sampling procedure outlined in Figure 4 was developed on the premise that research designs and educational research methods must serve both analytic and descriptive purposes. In an analytic sense, the use of the above sampling procedure for either within- or between-grade contrasts permits the independent influences of schooling and other CA-related factors to be estimated. However, the above sampling procedure has an added utility, that of permitting amount of schooling-performance functions to be generated in much the same manner that CA-performance functions are generated in research concerned with developmental phenomena.

That is, the use of the sampling procedure outlined in Figure 4 permits the cumulative influences of schooling to be estimated across grades. Such a schooling-performance function would be represented by adding the differences in performance for matched CA samples over different times of the school year for Ss in different grades; i.e., the estimate of the influences of schooling for the first educational period would be represented by: $\frac{\overline{X}_b - \overline{X}_a + (\overline{X}_d - \overline{X}_c)}{2}$ or by $\frac{\overline{X}_b + \overline{X}_d - (\overline{X}_a + \overline{X}_c)}{2}$. The estimate of the educational experiences for the second period of schooling would be represented by $\frac{\overline{X}_e + \overline{X}_g - (\overline{X}_d + \overline{X}_f)}{2}$.

The cumulative influences of schooling across educational periods and grades would be represented by pooling the estimates across these periods. This sampling procedure also permits CA-performance functions to be estimated independently of the influences of schooling. This would be accomplished by pooling performances differences for samples varying in CA who have equivalent educational experiences; e.g., $\frac{\overline{X}_c - \overline{X}_a + (\overline{X}_d - \overline{X}_b)}{2}$ or $\frac{\overline{X}_c + \overline{X}_d - (\overline{X}_a + \overline{X}_b)}{2}$.

Within- and Between-Grade Contrasts in Educational Research

It is emphasized that all educational problems and issues do not require a sampling plan as elaborate as that specified in Figure 4. In fact, most research problems probably require that only a section of the total sampling plan be used. Such a determination must be made by the individual researcher after taking into consideration the nature of the research problem, past empirical findings and the theoretical model or hypotheses to be investigated. However, it is of interest to note some of the additional phenomena which may be studied when within-grade contrasts and/or between grade contrasts are made in conjunction with the above sampling plan. For example, within-grade contrasts would be especially appropriate when the researcher is interested in cross-seasonal behavior changes in the children. For example, the amount of time spent in study may vary with the season of the year or the proximity to important holidays (e.g., Christmas). Similarily, between-grad contrasts for matched-CA samples at the end of one grade and the beginning of another may provide information concerning the (non-CA related) impact of changing school grades on children's behavior.

CA, ALTERNATE DEVELOPMENTAL SCALES AND RESEARCH METHODOLOGY

There has recently been considerable controversy and discussion concerning the role and use of CA in studies concerned with describing the nature and course of behavioral development (Baltes, 1968; Baltes & Goulet, 1971; Bijou, 1968; Birren, 1959, 1963; Goulet, 1970, 1973; Kessen, 1960; Neugarten, 1968, 1973; Neugarten & Datan, 1973; Schaie, 1965; Wohlwill, 1970, 1973). However, most of these papers have been concerned with the limitations of CA rather than considering the role(s) that it does play in developmental inquiry. Furthermore, the general concerns regarding the limitations of CA as a variable in developmental research are shared, but the reasons for this concern vary widely. The present sections represents an attempt to classify the various uses and limitations of CA from different theoretical perspectives, especially as they relate to attempts to identify developmental (as opposed to generation-related or secular change-related) changes in behavior.

Age Scales and Development

Kessen's (1960) statement defining the subject matter of developmental psychology provides an excellent base from which to describe the various uses of CA in developmental research. He proposed: "A characteristic is said to be developmental if it can be related to age in an orderly or lawful way," (p. 36). Apart from occasional and periodic reminders that age does not qualify as an experimental variable (e.g., Baltes, 1968), the functional statement R (response) = f (Age) has been generally accepted [even with its limitations (Birren, 1959; Wohlwill, 1973)], by most developmentalists as defining the subject matter of the field.

While not rejecting the importance of CA as an index of behavioral change, Neugarten and Datan (1973) suggest, "It is a truism that chronological age is at best only a rough indicator of an individual's position on any one of numerous physical or psychological dimensions. The significance of a given chronological age···when viewed from a sociological or anthropological perspective, is a direct funtion of the social definition of age." Similarly, Baer (1970) suggests that CA is used rather grossly as a cataloging device in order to manage the apparently unmanageable diversity and heterogeneity which exists among children. His comments highlight a number of important elements regarding the use of CA in developmental researchers. We suggest that the conventional methods of subject selection and matching in developmental research rarely consider the "point of origin" as a nominal property. Rather, the major concern is to describe and explain the behavior changes or differences which occur across time for selected populations. For example, researchers using Ss enrolled in school typically select and differentiate samples by school grade rather than chronological age. The CA range of the children within a specific school grade, however, typically meets or exceeds 12 months. Thus, even though the average difference in CA for Ss selected from successive grades will approximate 12 months (as the metric of time) the use of birth as a functional defining characteristic has been sacrificed.

Similar conventions exist in the literature concerning adult development and aging where the performance of Ss falling within specific CA ranges, e.g., 26-35, 36-45, 46-55, etc., are compared. Again, such a convention maintains equal time (or age) intervals between successive groups but scarifices the point of origin as one of the formal characteristics of a CA-based scale of development. In other words, the concern of the researcher has been to describe the developmental changes which occur across the time or age range included in the study using the developmentally "youngest" sample for comparison. One possible reason for this is that developmental and educational research does not, as yet, require a high degree of precision in matching variables (e.g., Baer, 1970). However, a central premise of this paper is that matching criteria are important since different uses of the point of origin serve as convenient cataloging devices to differentiate among various "types" of developmental research.

Three Uses of CA in Developmental Research

Wohlwill (1973), Baer (1970) and others suggest that CA, as an index along which to measure behavior change can be used as a purely descriptive (and thus causally neutral) scale. We suggest that such a position is appropriate only if the point of origin (e.g., birth) is disregarded as a functional characteristic in developmental inquiry. In other words, if time since birth is functionally irrelevant, then the only operative characteristic is the metric of time (in this case calendar time). However, a developmental scale must involve

both nominal characteristics, i.e., point or origin and metric of time. Chronological age is no exception. When CA is used as an index of development the investigator accepts birth by fiat as a significant life event against which to describe the course of behavioral development. Furthermore, birth, as a point of origin, specifies the manner in which Ss are to be matched or differentiated as to level of development.

A second use of chronological age by developmental researchers has been aptly discussed by Birren (1959) and Wohlwill (1973). Birren (1959) suggests that the aging process takes three forms; biological, psychological and social aging. Biological aging designates the position of the individual along his/her natural life span in ordinal units. Psychological aging refers to the achievements and potentials of the individual. Social aging refers to an individual's acquired social habits and status -- a composit of the individual's performance in social roles. Birren acknowledges the substantial degree of overlap between these three "types" of aging but suggests that these are the most likely candidates for alternate age scales. Since these scales currently do not exist, CA is used as a convenient substitute for underlying biological, psychological or sociological processes and is assumed to correlate with each of them. Given that CA is used as a measure reflecting some underlying process, several assumptions have to be made: first, the "point of origin" of the process must be correlated with birth, and; second, a linear relation exists between the underlying process and CA at least over the ages or period of interest.

The third form of a CA scale may be designated as a state or stage scale. Such a scale may take different forms, but the defining characteristic is that a particular period within the life-span of an individual is charted by points (designated by CA) of transition from one developmental status to another. State-oriented scales are similar to process-oriented scales discussed above in that the theoretical basis of such a scale may have biological, sociological, or psychological underpinnings. The major difference between the two types of scales is that state- or stage-oriented developmental scales assume at least some degree of discontinuity of processes between adjacent developmental periods.

Neugarten and Datan (1973) point out that, "Although anthropologists···have pointed to discontinuities in cultural conditioning at various points in the life cycle, the recognition of the need for resocialization in adulthood is relatively new." They suggest that "new learning" across the life span occurs in respons to, or anticipation of, the succession of life tasks (or social roles) which individuals adopt. For example, familiar "transition" points on a sociological scale are entry into school, marriage, retirement, etc. The criterion for selecting important transition points is that the

social role in question be accompanied by a relatively circumscribed set of behavioral expectations. In this regard, there is strong agreement among members of a society concerning the salutatory significance of life events (Neugarten & Datan, 1973).

Discontinuous state scales have been developed from a psychological and biological perspective. For example, the major periods in Piaget's theory (e.g., sensory-motor, preoperational, concrete operations, and formal operations) constitute fundamentally discontinuous stages in the individual's life span and describe a specific set of behaviors. Similarly, puberty constitutes a biologically related transition period.

The use of CA to mark transitions between stages requires that CA and the succession of social, psychological, or physical states be highly correlated. Neugarten and Datan (1973) have provided such evidence from a sociological perspective by noting a high degree of consensus regarding the timing (in terms of CA) of major life events in an individual's life span. Similarly, there is general agreement among diverse sets of respondents regarding the chronological age boundaries differentiating life periods, (e.g., English and English, 1957; Neugarten, Moore, and Lowe, 1956).

Reconsideration of the Longitudinal Method and Behavioral Development

The study of developmental changes in behavior spawns a single, basic research paradigm -- the longitudinal method. The defining property of the method is that a single individual is tested at two or more points in time. It is also important to note that the method is theoretically neutral since its use does not require the investigator to adopt a specific developmental scale along which to chart the sequence of human development. If longitudinal measurements were collected for several individuals the resultant data permit conclusions to be drawn regarding the interindividual similarities in the sequence of behavioral development. When marked similarities in the sequence of occurrence of behaviors are observed among the individuals studied, the regularities cannot be charted on a developmental scale since the longitudinal method makes no reference either to the point of origin or the metric of change. The developmental scale adopted for this purpose should be the one which is most highly correlated with the behavior studied. Once adopted the scale specifies the manner in which the data of individual Ss are to be grouped and the nature of the time intervals across which the behaviors are to be described.

Therefore, alternative developmental research methods are derivable only after the investigator adopts a theoretically meaningful scale.[1] For example, cross-sectional measurement is often used as a convenient substitute

[1] In this paper, the subsequent use of "developmental scale" is to be taken in the above described genetic sense and not in reference to any specific metric.

for longitudinal measurement. The selection of the different groups of Ss for testing requires that the researcher choose a specific developmental scale. Once the scale is chosen, the criterion for subject selection and matching become apparent. Additionally, it is now possible to specify the alternate longitudinal and cross-sectional design specified by the scale.

In short, the longitudinal method is a theoretically neutral and generalized research method in developmental inquiry. Furthermore, when used in its generalized form, it provides data concerning the sequence but not the temporal course of behavioral development. Special cases of the longitudinal method (along with their cross-sectional counterparts) are derivable only when the researcher adopts a developmental scale. For example, if CA is selected as the scalar metric, Ss are matched or differentiated according to CA and can therefore be selected and tested according to either longitudinal or cross-sectional sampling procedures.

Each developmental scale spawns its own unique longitudinal method. A process-oriented developmental scale, for example, may involve selecting and matching Ss according to a biological, sociological, or psychological process (e.g., skeletal age, Shuttleworth, 1937) and testing the Ss at selected points in time (defined by either calendar units or process-related criteria) thereafter. Similarly, stage- or state-scales of behavioral development would specify matching criteria defined by the stages or states in question. Neugarten and Datan (1973), for example, have described an alternate longitudinal paradigm in which the point of origin differs from a CA-based scale but which retains the same metric of time. In this regard, the functional point of origin of a particular behavioral sequence may be the acceptance of a particular social role (e.g., fatherhood) and the patterns of behavior change following this event can be charted on a scale of calendar time, e.g., fatherhood, fatherhood + one unit, fatherhood + two units, etc.

The striking parallels between CA-based and process-oriented scales are readily apparent. In both cases, behavior change is charted in terms of proximity (measured in units of calendar time) to an important life event. In addition, birth (or a descriptive CA-based scale) and fatherhood (on a process-oriented sociological scale) provide the only "benchmark" or point of origin. This suggests an underlying continuity of behavior change across time marked from the point of origin of the behavior being studied. The scales differ, however, since Ss are matched (and differentiated) according to criteria defined by the different "functional" points of origin for the two scales.

Parallels to the longitudinal paradigm proposed by Neugarten and Datan (1973) also exist utilizing theories focused on biological/psychological processes. As an example, the classic study by

Shuttlesworth (1937) provided data concerning the correlation between puberty and the "growth spurt" in adolescence. This was accomplished by matching Ss for the onset of puberty (rather than CA) and charting physical growth from this point forward. Within a psychological framework, Piaget (e.g., 1928) also accepts this method by suggesting that the sequence of behavior change follows a universal order starting with the onset of psychological periods and stages. Interestingly, Bijou and Baer (1961, 1965) follow a very similar line of reasoning to that of Neugarten and Datan (1973) by suggesting that environmental "setting events" influence behavior throughout life.

The preceding discussion has highlighted several important points related to subject selection and matching in developmental research. First and foremost, the adoption and use of a specific developmental scale requires the researcher to adopt certain assumptions relating to point of origin and the metric of time. However, as has been suggested, the nominal properties of the point of origin are rarely considered in developmental research. Rather, the concern in most research is with the study of a developmental process and how it changes with time. Subjects are chosen and tested on the basis of representing the ages or time periods over which the process is thought to change. In such cases, the functional point of origin for the developmental study in question is the developmentally "youngest" sample. In such cases the nominal and functional point of origin for the researcher may be different, e.g., birth vs. six-year-olds; yet the nominal and functional metric of time may be identical (e.g., units of calendar time such as months, years, etc.).

It is important at this point to discuss additional limitations of the sampling model proposed by Schaie (1965). First, Schaie limited his model to situations where the researcher has adopted a CA-based scale of behavioral development. This is an unnecessary restriction of the model. In addition, two additional limitations of the model are at issue here.

The first limitation discussed earlier, has received considerable attention by others (e.g., Baltes, 1968; Baltes & Nesselroade, 1974; Buss, 1973, Schaie, 1965; Wohlwill, 1973) concerns the functional independence of the components of age, cohort, and time. For example, Baltes' (1968) suggestion that the three components are not mtually independent, i.e., once two components have been defined, the third is fixed, is relevant here. As Buss (1973) and Wohlwill (1973) have argued, such criticisms relate to methodological rather than theoretical concerns. Even though any two of the components cannot be functionally varied independently of the third, the concepts of developmental (age) generational (cohort), and secular (time-related) change to indeed qualify as separate theoretical concepts (e.g., Buss, in press; Troll, 1973).

It is important to highlight two additional aspects of the issue concerning the independence of the three components. The first aspect concerns the manner in which the three components are defined and the way in which populations are matched. First, Schaie's model, by adopting CA as a developmental scale not only restricts the researcher to indexing behavioral development from birth as a point of origin, but also confines the definition of cohort to data of birth rather than some alternate definition, such as, the population of children who entered first grade in September, 1975, etc.

Any deviation from a CA-based scale requires modification of the general model proposed by Schaie (1965). As an example, if subjects to be tested were in terms of a sociological state (as a level of development) and time of testing, e.g., all subjects who were married for the first time in September 1975, the third component, cohort, would lose all functional meaning when defined in terms of birthdate. Similarly, if cohort is defined in terms of "family lineage" or one of the alternate accepted definitions of generations and generational change (e.g., Troll, 1973), time of measurement may be specified, but CA loses theoretical and functional meaning. The point is, if a developmental scale other than a CA-based one is selected for use, all three components must be re-examined both methodologically and theoretically.

The second limitation of Schaie's developmental model concerns the restrictive manner in which the second formal characteristic of time-related scales (the metric of change) is defined. That is, the use of Schaie's model restricts the investigator to a scale of calendar time rather than one which might more properly fit the phenomenon under study. While it would be possible, for example, to identify samples of subjects on a scale of biological development (e.g., skeletal age) and to the samples at selected testing points (e.g., September 1975, and September, 1976) the second testing point would have to occur after an equal time interval for all subjects or else the functional meaning of time of measurement (as defined by Schaie) would be lost. In addition, even though the above research design (skeletal age x time) conforms in some respects to Schaie's (1965) cross-sequential design, the main effects of time of measurement would more properly reflect developmental change than secular change for the two populations.

The above discussion is not meant to discount the importance of the concepts of age, cohort, and time of measurement in the study of behavioral development. Indeed, the present analysis reaffirms the need to incorporate variants of Schaie's sequential analyses as necessary paradigms in developmental research. In fact, the present analysis suggests two additional types of variants of Schaie's sequential paradigms, and leads to the conclusion that Schaie's model itself is restricted in its generalizability. These points are discussed in Chapter 3.

REFERENCES

Baer, D. M. An age-irrelevant concept of development. Merrill-Palmer Quarterly, 1970, 16, 238-245.

Baltes, P. B. Longitudinal and cross-sectional sequences in the study of age and generation effects. Human Development, 1968, 11, 145-171 (a).

Baltes, P. B., & Goulet, L. R. Exploration of developmental variables by manipulation and simulation of age differences in behavior. Human Development, 1971, 14, 149-170.

Baltes, P. B. & Nesselroade, J. R. Cultural change and adolescent personality development: An application of longitudinal sequences. Developmental Psychology, 1972, 7, 244-256.

Bell, R. Q. Convergence: An accelerated longitudinal approach. Child Development, 1953, 24, 145-152.

Bijou, S. W. Ages, stages and the naturalization of human development. American Psychologist, 1968, 23, 419-427.

Bijou, S. W. & Baer, D. M. Child Development, Vol. 1. A systematic and empirical theory. New York: Appleton, 1961.

Bijou, S. W. & Baer, D. M. Child Development, Vol. 2, New York: Appleton, 1965.

Birren, J. E. Principles of research on aging. In J. E. Birren (Ed.), Handbook of aging and the individual. Chicago: University of Chicago Press, 1959.

Birren, J. E. The psychology of aging. New York: Prentice-Hall, 1964.

Buss, A. R. An extension of developmental models that separate ontogenetic changes and cohort differences. Psychological Bulletin, 1973, 80, 466-479.

Buss, A. R. Generational analysis: Description, explanation, and theory. Journal of Social Issues, in press.

English, H. B. Chronological divisions of the life span. Journal of Educational Psychology, 1957, 48, 437-439.

Goulet, L. R. Training, transfer, and the development of complex behavior. Human Development, 1970, 13(4), 213-240.

Goulet, L. R. The interfaces of acquisition: Models and methods for studying the active, developing organism. In J. R. Nesselroade and H. W. Reese (Eds.), Life-Span Developmental Psychology: Methodological Issues. New York: Academic Press, 1973.

Goulet, L. R., Hay, C. M. & Barclay, C. R. Sequential analyses and developmental research methods: Descriptions of cyclical phenomena. Psychological Bulletin, 1974.

Goulet, L. R., Williams, K. G. & Hay, C. M. Longitudinal changes in intellectual functioning in pre-school children: Schooling and age-related effects. Journal of Educational Psychology, 1974.

Kessen, W. Research design in the study of developmental problems. In P. H. Mussen (Ed.), Handbook of research methods in child development. New York: Wiley, 1960, 36-70.

Neugarten, B. L. Adult personality: Toward a psychology of the life cycle. In Neugarten, B. L. (Ed.), Middle Age and Aging, Chicago: University of Chicago Press, 1968, 137-147.

Neugarten, B. L. Personality change in late life: A developmental perspective. In C. Eisdorfer & M. P. Lawton (Eds.) The psychology of adult development and aging. Washington, D. C.: American Psychological Association, 1973, 311-338.

Neugarten, B. L. & Datan, N. Sociological perspectives on the life cycle. In P. B. Baltes and K. W. Schaie (Eds.), Life-span developmental psychology: Personality and socialization. New York: Academic Press, 1973, 53-71.

Neugarten, B. L. Moore, J. W. & Lowe, J. C. Age norms, age constraints, and adult socialization. American Journal of Sociology, 1965, 70, 710-717.

Schaie, K. W. A general model for the study of developmental problems. Psychological Bulletin, 1965, 64, 92-107.

Shuttleworth, F. K. Sexual maturation and the physical growth of girls age six to nineteen. Monographs of the Society for Research in Child Development, 1937, 2, No. 5.

Troll, L. E. Issues in the study of generations. Aging and Human Development, 1970, 1, 199-218.

Wohlwill, J. F. The age variable in psychological research. Psychological Review, 1970, 77, 49-64. (b)

Wohlwill, J. The study of behavioral development. New York: Academic Press, 1973.

CHAPTER 3

GENERAL SAMPLING STRATEGIES FOR B = f(T) RESEARCH

General Sampling Designs for B = f(T) Research

In Chapter 2, the discussion highlighted the fact that Schaie's general developmental model represents only one of a family of sampling strategies amenable to the study of behavior changes associated with time. Other models, similar in form to the one Schaie (1965) proposes, may be derived whenever the researcher adopts a developmental scale other than CA.

The first variant of Schaie's (1965) sequential analyses parallel his general developmental model with the exception that a developmental scale other than CA is used.' Figure 1 provides an example of the model using a developmental index based on sociological criteria. Samples of Ss (cohorts) who were married for the first time in 1970, 1975, and 1980 are tested at the time of marriage and in increments of five years thereafter.

The use of Schaie's developmental model requires that the age and cohort variables share the same nominal and/or functional point of origin. The choice of a sociological scale of development (time since marriage) leads to a redefinition of the cohort variable (year of marriage) in the same manner that CA as a developmental index presupposes a definition of cohort based on date of birth. Nevertheless, a sampling design such as that provided in Figure 1 permits cohort-sequential, time-sequential, and cross-sequential analyses to be performed if a minimum of six samples of Ss conforming to the sampling design in Figure 1 are represented.

Figure 1 provides an example of an alternate model based on sociological criteria and parallel models may be derived using psychological or biological criteria.

The paradigms basically conform to Schaie's model, and share some of the same attributes and limitations. The attributes have been fully documented by Schaie (1965), Baltes (1968) and in the present paper. The major limitation of Schaie's (1965) model is that the three components of developmental change (age, cohort, and time-of-testing) cannot be defined independently of one another and this limitation is shared by the variant of the general model presented in Figure 1. As was mentioned in Chapter 2, such difficulties arise when the scales used to define the age and cohort variable share the same nominal and/or functional point of origin.

However, it is possible to generate sequential paradigms analogous to time-, cohort-, or cross-sequential sampling strategies which

Time of Measurement	Age Level A_1: M	A_2: M + 5 years	A_3: M + 10 years
1970	1970[a]		
1975	1975[b]	1970[c]	
1980	1980[d]	1975[e]	1970[f]

*Cell entries refer to cohort groups defined by rate of marriage (M)

Figure 1

A Sampling Model for Developmental Research

Based on Sociological Criteria*

do not share this limitation. Figure 2 provides one example of a variant of a cohort-sequential design. Cohort is defined by family lineage and developmental level by the sociological state of marriage.

The second variant of Schaie's sequential analyses is derivable if the assumption is made that age (maturation), cohort (generation) and time (secular change) are defined independently of one another.

The research paradigms parallel the sequential designs proposed by Schaie in that generational, secular, and age changes are the focus of the investigation. The paradigms also adopt calendar time as the metric. However, since the components of age, cohort, and time-of-measurement are by definition uncorrelated, the paradigms differ from those proposed by Schaie (1965).

CA and Other Age Scales of Development

The previous discussion has highlighted the similarities between CA- and alternate developmental scales. It was shown that each scale generates its own prototype of longitudinal and cross-sectional sampling strategies and its own variant of the sequential strategies proposed by Schaie (1965).

The final type of design to be proposed here examines the relationships between CA-, sociological-, biological-, and/or psychological-scale(s) of development.

Such investigations could take the form specified in Figure 3a, where Ss representing different levels of CA are tested at the point of marriage and five years thereafter. The differences between the row means represent effects attributable to CA, whereas differences between the column means reflect effects which covary with time since marriage. Both "independent" variables are developmental in nature and the results from such an investigation permit inferences to be made regarding the degree to which performance varies with CA, time since marriage, or both. And, as such, the design provides information regarding the sensitivity of two alternate age-scales to the phenomenon of interest. Nevertheless, the design, even though calendar time of measurement is controlled as with any cross-sectional sampling procedure does not permit the cohort influences to be separated from those related to development.

Figure 3b represents another variant of such a design. It conforms in some respects to Schaie's time-sequential design in that CA and time of testing are factorially varied. However, in this case, both CA and time of testing are factorially varied. However, in this case both CA and time since marriage correlate perfectly with calendar time (1970, 1975), i.e., Ss from both cohorts were married in 1970.

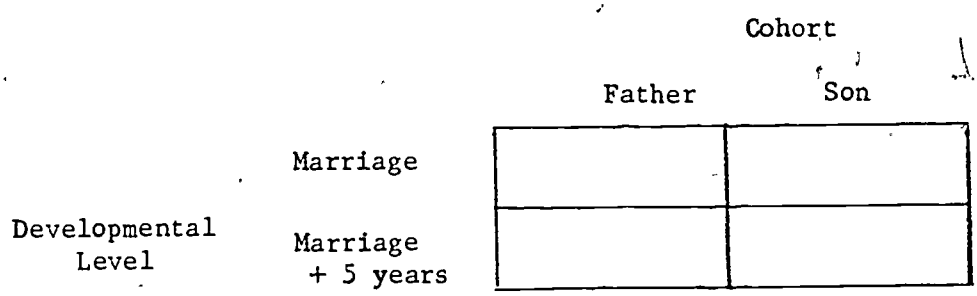


Figure 2

A Cohort-Sequential Design Based on Independently-Defined Cohort and Age Levels

Figure 3a

CA at time of testing	Marriage	M + 5 years
25	1970*	1975
30	1975	1980

Figure 3b

CA at time of testing	Marriage	M + 5 years
25	1970*	1975
30	1970	1975

*Cell entries correspond to times of measurement

Figure 3

Sampling Designs for Developmental Research Varying Developmental Level Along Two Dimensions*

The merit of designs such as those described in Figures 3a and 3b from the framework of an educational perspective is best illustrated by a reconsideration of the sampling model for educational research presented in Chapter 2 (Figure 4) and presented in another form in Figure 4.

There is a paucity of data available utilizing the paradigm exemplified in Figure 4. However, scrutiny of literature reveals a set of studies (Baltes & Reinert, 1969; Schaie, 1972) conducted for other purposes but which nevertheless provide for comparisons in which CA and amount of exposure to school curricula are orthogonally varied. Furthermore, there are several sets of data emanating from our laboratory which were conducted for the primary purpose of testing the utility of the sampling procedures presented in Figure 4. These data provide for within-grade contrasts (Goulet, Williams, & Hay, 1973, in press; Goulet, Williams, Bozinou & Hexner, 1973; Wood & Goulet, 1973a), and between-grade contrasts (Wood & Goulet, 1973).

In view of the recent availability of such data, it is considered important to present the results in summary form and to discuss the studies themselves in considerable detail. The studies provide information regarding the independent behavioral correlates of schooling and CA for children across the range of CA from four to nine years and from nursery school to fourth grade. Also, data are available across a variety of behavioral domains including intellectual growth (Baltes & Reinert, 1969; Goulet, Williams & Hay, 1974; Schaie, 1972) visual-perceptual performance (Wood & Goulet, 1973a, 1973b) for single-trial free recall performance, subjective estimates of recall ability (Goulet, Williams, & Hay, 1973), and the utilization of rules of addition (Goulet, Williams, Bozinou & Hexner, 1973).

Summaries of each of the sets of data providing within-grade contrasts are presented in Table 1 and are identified by author and the available measure of performance. Table 2 provides the data from the single study (Wood & Goulet, 1973b) where between-grade contrasts are possible. In each instance except where noted, CA and time of testing noted, CA and time of testing in the school year are varied and superior performance is reflected by higher scores. The row and column means for each of the matrices in Table 1 represent performance for the main effects of Time of Testing and CA, respectively. In each case, the data represent means based on independent samples and the data are amenable to analysis within a 2 x 2 factorial design with CA and Time of Testing as the two factors. In addition, with the exception of parts of the Baltes and Reinert (1969) data or where noted, the main effects for CA and for Time of Testing are statistically significant. No interactions were evident in the data.

In each case the data represent the performance of children who were enrolled in the appropriate grade for their age. To eliminate the possibility of a selection bias related to grade placement, the

	School Grade 1					
	1			2		
Month of Testing	Sept.	Jan.	May	Sept.	Jan.	May
			6-6[c]	6-10[f]	7-2[i]	7-6[k]
Chronological Age at Time of Testing		6-6[a]	6-10[d]	7-2[g]	7-6[j]	
	6-6	6-10[b]	7-2[e]	7-6[h]		

Figure 4

An Extended Sampling Strategy for

Testing School-Age Children

Table 1

Summary Means for Research Permitting Within-grade Variation of CA and Amount of Schooling

Baltes-Reinert (1969)

Letter Series

		CA 8-4	CA 8-8	$\bar{X}$	CA 9-4	CA 9-8	$\bar{X}$
TT	March	11.8	12.0	11.9	13.8	14.1	13.9
	July	12.2	12.5	12.3	14.2	14.5	14.3
	$\bar{X}$	12.0	12.3		14.0	14.3	

Word Completion

		CA 8-4	CA 8-8	$\bar{X}$	CA 9-4	CA 9-8	$\bar{X}$
TT	March	5.7	6.9	6.3	9.8	9.6	9.7
	July	6.2	7.7	6.9	9.7	10.4	10.1
	$\bar{X}$	5.9	7.3		9.7	10.0	

Table 1 (Continued)

Basic Arithmetic

TT		CA 8-4	CA 8-8	$\bar{X}$	CA 9-4	CA 9-8	$\bar{X}$
	March	16.4	18.0	17.2	22.1	21.5	21.8
	July	17.1	18.3	17.7	21.2	21.8	21.5
	$\bar{X}$	16.7	18.1		21.7	21.6	

Letter Counting

		8-4	8-8	$\bar{X}$	9-4	9-8	$\bar{X}$
	March	33.9	35.5	34.7	38.5	38.7	38.6
	July	31.3	32.6	31.5	38.1	38.6	38.3
	$\bar{X}$	32.6	34.1		38.3	38.7	

Table 1 (Continued)

Schaie (1972)

Mental Age

		Boys CA 6-6	Boys CA 6-10	$\bar{X}$		Girls CA 6-6	Girls CA 6-10	$\bar{X}$
TT	Fall	72.8	82.0	77.4	Fall	78.9	77.8	78.3
	Winter	84.2	91.5	87.9	Winter	81.3	92.7	87.0
	$\bar{X}$	78.5	86.7		$\bar{X}$	80.1	85.3	

Goulet-Williams-Hay (1974)

		CA 4-4	CA 4-9	$\bar{X}$
	Oct.	66.9	69.3	68.1
	March	69.3	78.1	73.7
	$\bar{X}$	68.1	73.7	

Goulet-Williams-Bozinou-Hexner (1973)

Errors to Criterion

Treatments

		Rule CA 6-5	Rule CA 6-10	$\bar{X}$		Interference CA 6-5	Interference CA 6-10	$\bar{X}$
TT	Nov.	18.0	22.8	20.4	Nov.	29.8	24.4	27.1
	April	10.2	7.1	8.6	April	37.5	29.1	33.3
	$\bar{X}$	14.1	15.0		$\bar{X}$	33.7	26.7	

Table 1 (Continued)

Goulet-Williams-Hay (1973)

		Estimation				Recall Span		
		CA				CA		
		4-4	4-9	$\bar{X}$		4-4	4-9	$\bar{X}$
TT	Nov.	8.5	9.1	8.8	Fall	10.2	9.1	9.7
	April	7.1	8.0	7.5	Winter	6.0	7.7	6.9
	$\bar{X}$	7.8	8.5		$\bar{X}$	8.1	8.4	

Wood-Goulet (1973a)

		Errors		
		CA		
		5-4	5-10	$\bar{X}$
TT	Oct.	12.0	13.9	13.0
	April	9.6	9.7	9.6
	$\bar{X}$	10.8	11.8	

Table 2

Summary Means for Research Permitting

Between-Grade (Matched-CA) Contrasts

Wood-Goulet (1973b)

Errors

		Grade K (5-10)	Grade 1 (5-11)	$\bar{X}$
TT	Oct.	13.8	9.5	11.7
	April	7.6	6.5	7.0
	$\bar{X}$	10.7	8.0	

the children were selected for testing from the middle 70 percent of the age range within a class; i.e., the youngest and oldest children within a grade were not sampled.

Table 1 provides data taken from Baltes & Reinert (1969). The data represent raw score performance on each of four subtests of intelligence (including letter series, word completion, basic arithmetic, and letter counting) which were collected in the months of March and July for samples ranging in CA from 8-4 to 8-8 (third grade) years in Study I, and 9-4 to 9-8 years (fourth grade) in Study II. Therefore, only the directionality of results is discussed. As is apparent, the diagonal contrast (upper-left and lower-right cell means) provides data representing longitudinal changes in performance, the vertical (cross-row) contrast represents a time-lag comparison, and the horizontal (cross-solumn) contrast represents a cross-sectional comparison. Only the longitudinal comparison involves mean differences which confound CA and length of schooling. As may be seen from these data, the longitudinal contrasts provide an estimate of change which exceeds that of the cross-sectional and time-lag contrasts. Also, with the exception of the letter-counting measure, the column and row means suggest that amount of school experience and CA are each positively correlated with performance. With the letter-counting measure, the relationbetween CA and performance is positive and the relation between amount of school experience and performance is negative. Such opposing effects of the two variables leave a longitudinal function which suggests no (or even slightly negative) changes in performance over the four-month interval which separated the two testing periods.

The second sets of data in Table 1 are taken from studies by Schaie (1972) and Goulet, Williams and Hay, 1974. The cell means represent the Mental Age of first-grade (Schaie, 1972) and nursery-school children (Goulet, Williams, & Hay, 1974). Intellectual performance was found to relate positively to amount of schooling and to CA for both samples ofr measures which were taken in 1933 (Schaie, 1972), and 1973 (Goulet, Williams, & Hay, 1974) and for both boys and girls (Schaie, 1972).

The third set of data were taken from Goulet, Williams, Bozinou, and Hexner (1973). The cell means represent performance on a paired-associates transfer task. In the Rule condition, rapid acquisition was expected if the children (first-grade) used an addition rule of "add 1" to learn the individual paired associates in the list. Nonuse of the rule would interfere with performance. Thus, superior performance is reflected by fewer errors to criterion. In the Interference condition, the children learned a transfer list of paired associates where no rule was possible and interference (negative transfer) was expected. As the data suggest, superior performance was positively

related to amount of schooling in the Rule condition, whereas the reverse was true in the Interference condition. Chronological age was unrelated to performance in the Rule condition, and the older children learned the transfer task faster (fewer errors) in the Interference condition.

The data provided by Goulet, Williams, and Hay (1973) take two forms. The first set of data refer to childrens' estimates of their ability for immediate recall. The children were shown up to 10 familiar, but unrelated, pictures and they were asked to judge how many they could remember if they were shown once. The second set of data refers to the childrens' actual recall span; i.e., the longest series of pictures they could remember without error after one presentation. As may be seen from these data, subjective estimates of recall ability relate positively to CA and negatively to amount of schooling. For the data on recall span, null effects of CA and negative effects related to amount of schooling are found.

The data taken from Wood and Goulet (1973a) represent raw score performance on the Bender-Gestalt Visual Motor Test. The data represent error scores so superior performance is represented by lower scores. Again, amount of schooling is positively related to better performance, with null effects related to CA.

The last set of data (presented in Table 2) deviate substantially from those contained in Table 1. First, the data provide for between-grade contrasts of matched-CA children. Second, the data provide for longitudinal measurement for these samples across the period from October to April. Thus, the main effect related to school grade represents performance differences for samples who differ by one year in amount of schooling. The main effect for time of measurement, as with all longitudinal contrasts confounds CA and time of testing and thus the results cannot be unequivocally attributed to factors related to CA or schooling. Nevertheless, the between-grade effect suggests pronounced facilitative influences of schooling even though the Ss are matched on CA.

Data such as those presented in Tables 1 and 2 provide support for the utility of utilizing sequential sampling strategies when age (developmental level) is varied simultaneously with two developmental scales.

There are, a number of issues which warrant further consideration. The first point of concern is that most small-scale studies and certainly all available large-scale studies of student development have relied on simple cross-sectional or longitudinal sampling procedures. Examples here are the Survey of Equality of Educational Opportunity (Coleman, 1971) which used a cross-sectional design and the Growth Study conducted at the Educational Testing Service (Anderson & Maier, 1963; Hilton & Meyers, 1967) which involved a longitudinal design. As Hilton and Patrick (1970) have noted, the results of both of these studies confound the developmental changes of primary interest with

generational or secular change factors, respectively, which occurred for the samples tested. Just as important for present purposes, the above studies were initiated for the purpose of explicating the influences of school experiences across grades and yet provide no estimates of these effects.

The data provided in Tables 1 and 2 uniformly provide support for the assumption that influences of schooling exist independently of those which may be expected from normal aging; i.e., from the cumulative influences of past experience and/or maturation (Baltes & Goulet, 1971; Schaie, 1965), and also suggest the utility of providing independent estimates of performance associated with nonschool-related changes in chronological age. Such estimates become especially important under conditions where the factors associated with CA and school experience may have opposing effects, (e.g., Baltes & Reinert, 1969; Goulet, Williams, Bozinou, & Hexner, 1973; Goulet, Williams & Hay, 1973). In this regard, the suggestions offered here parallel those of Schaie (1965), Baltes (1968), Hilton and Patrick (1970) and others who have been primarily concerned with separating sources of variance associated with generational, secular, and age change in student development.

Nevertheless, it is not the intent here to elevate either chronological age nor amount of school experience to the status of an experimental/independent variable. Chronological age remains a descriptive, biotic variable (as indeed does school experience in the context in which it is used here) since it cannot be experimentally manipulated, nor replicated. That is not to say that CA is a useless variable. It remains one of the most useful ways in which to classify or categorize children, (Baltes & Goulet, 1971; Kessen, 1961; Wohlwill, 1970) and by which to chart behavioral change in research of a developmental nature. In the context of the present paper, CA-related changes in behavior are divided into two components, those which vary with schooling, and those associated with nonschool-related changes associated with CA.

A second point is that none of the problems in educational research are vitiated by the use of school grade, rather than chronological age, in such studies. Such distinction is obviously important in educational research but only to the extent that it is made meaningful through the assessment of the behavioral changes which occur over the school year for the grade samples tested and to the extent that other CA-related factors are controlled.

It is also important to mention that the sampling strategy suggested in Figure 4 is similar to certain popular designs used in educational research. One example is the time by treatment design where two or more randomly selected groups of children matched in CA, school grade, etc., are exposed to different school curricula over some instructional period and the performance of the groups is

contrasted at the end of the instructional period. Such a design, which involves elements of both longitudinal and experimental methods, controls for CA between the two groups of children. Unfortunately, the design suffers from the fact that the children are both older and have undergone the instructional sequence at the end of training. Thus, the performance differences among the experimental groups reflects not only the independent influences of the instructional sequence but also the interaction between CA and the instructional treatments in influencing performances (Goulet, 1970). This inference holds even though Campbell and Stanley (1963) refer to such a design as a "true experimental design." It is not until CA is incorporated into the design that the interaction of CA and instructional treatments and the independent influences of the instructional treatment upon performance may be separated. As is apparent, this modification of the design has each of the elements of the sampling plan exemplified in Figure 2 -- of course, with the desirable addition of an experimental treatment.

The primary issue considered in this paper concerns the assessment of the effects of educational intervention (used in the broad sense) on performance over the period of a school year or shorter interval. However, as has already been mentioned, the influences of schooling are usually not discernible from other CA-related influences on performance. That is not to say that the impact of or effects of exposure to the school curriculum can be considered to be independent of behavioral development. Rather, school learning must be considered to be one of the components in the developmental process. It is for the latter reason that alternate experimental designs have been developed in developmental psychology to provide estimates of the effects of educational experiences on performance unbiased by behavioral development. One such design involves the simulation or "acceleration" of the process through the provision of massed training or practice (Baltes & Goulet, 1971; Goulet, 1968). Such an experimental strategy is used very often in contemporary studies concerned with cognitive development (e.g., Sigel & Hooper, 1968; Gellman, 1969). However, such approaches, although appropriate for the study of developmental phenomena, cannot be generalized directly to school situations. This is true because: (1) It is not possible either to identify the range of experiences acquired in or as a direct result of the interaction in school; nor is it possible to simulate them in their entirety in controlled or laboratory situations; and, (2) Behavioral change induced through massed practice over a short term must, of necessity, be limited in scope. Also, attempts to generalize the findings to school situations are severely limited because of the possibility of an interaction between time and the acquisition of the behavioral phenomena of interest. In other words, the product of school experiences are acquired over a long period and through a variety of media, including the teacher, age-mates, and non-school situations prompted by school curriculum. There is no reason to expect that the effects of massed practice on specified tasks have effects which are isomorphic with those which are acquired as a result of schooling over the school year. Finally, studies using such a design focus (implicitly or explicitly)

on the identification of variables which influence student learning rather than on the description of education-related behavior change. While such research is needed, it does not lead to the types of information provided when using the sampling plan suggested here.

There is a second way to provide direct estimates of the effects of school experience which are unbiased by independent time or age-related components of behavioral change. In the most simple case, the procedure would involve the comparison of two groups of children across time (e.g., the school year) under conditions where both groups were eligible for acceptance into school but where one of the two groups was enrolled in school and one wasn't. However, it is extremely difficult to find "random" samples of children who are of school age but who have not been enrolled in school. And, even if such a sample were available in the general population it would be impossible to match them with children who were enrolled. The very conditions which precipitated the lack of enrollment would bias the sample. Campbell and Stanley have discussed these issues in detail. As is apparent, the sampling plan presented in Figure 2 utilizes a research strategy which capitalizes on the latter method while avoiding the potential sources of confounding when it is used.

SCHOOL EXPERIENCES, CA, AND THE DIRECTIONALITY OF BEHAVIOR CHANGE

The intent of this paper is not to comment directly on either the nature of the influences of schooling or the relation between performance and amount of schooling. Nor is it possible to specify a priori within the context of the sampling plan exemplified in Figure 2, either the magnitude or direction of the influences of factors related to CA and school experience on performance. Nevertheless, it is appropriate at this time to reiterate some of the general inferences which may be drawn from the data presented in Tables 1 and 2 and other sections of the paper. These inferences are provided below and appropriate discussion follows each point.

1. Available data suggest the utility of adopting the sampling plan in Figure 4 for educational research purposes and, although only few available studies permit contrasts of the type required, each provides evidence suggesting independent effects associated with CA and amount of schooling over periods as short as four months.

2. The relation between CA and performance and amount of schooling and performance may be complementary (either positive or negative) or opposing over the same period.

The point of interest here is that the relation between CA and performance is not uniformly positive during the years of formal education. In fact there is a substantial amount of evidence suggesting, for example, that the relation between CA and performance in problem-solving tasks is curvilinear over the age range from three to eighteen (e.g., Goulet & Goodwin, 1970; Weir, 1964). While the series of studies from which such inferences were drawn have involved cross-sectional sampling

procedures, there are probably many instances of behaviors which correlate positively with CA and negatively with amount of schooling (or vice versa) over the same time period.

3. A basic premise here is that designs used in educational research require sampling and testing at least at two points within the school year for Ss in the same grade. It is only with such a sampling plan that the behavior changes which occur over this period can be assessed. Such suggestions have already been made (e.g., Campbell & Stanley, 1963) and further reiteration regarding this point is unnecessary. Nevertheless, within-year as opposed to between-year times of testing should also minimize confounding due to attrition in educational research (e.g., Hilton & Patrick, 1970).

4. A central assumption is that the non-school related correlates of behavioral development (as indexed by variations in CA) must be controlled before the influences of educational intervention can be assessed. This assumption is similar to that made by Schaie (1965) and Baltes (1968) in their attempts to differentiate age change from generational and secular change in developmental research.

5. Although measures of achievement over periods of schooling generally show at least modest gains, reviewers of such research have been quick to mention that the achievement gains observed are as likely attributable to "maturation" as to the influences of instruction (Austin, Rogers, & Walbesser, 1972). Furthermore, such reviewers have lamented the fact that educational research directed to assessing the influences of schooling have provided no data demonstrating that the gains were maintained over time, especially in contrast to groups not exposed to instruction over the same period. The use of the sampling plan in Figure 4 provides for such estimates.

6. The suggestions contained in the present paper also hold in the context of the norming and standardization of achievement tests. That is, most standardized tests have utilized either cross-sectional or longitudinal sampling procedures in obtaining their normative sample. The biases which result from such a sampling procedure will vary as a result of date of testing, type of sampling procedure used, and the relation between amount of schooling, CA and performance on the standardized test. These biases have been demonstrated by Goulet, Williams and Hay (1974) and readers are referred to this paper for a complete discussion of this point.

Some final comments concerning the influences of schooling are warranted. First, there is no intent to imply that the results attributed to the influences of school experience in the present study are directly or exclusively attributable to the "in-classroom" experiences of the children. Rather, such influences may take many forms, ranging from the effects of the different forms of social interactions,

environmental contexts, and parental or peer demands which confront the children while they are enrolled in school. Such potential caveats do not vitiate the use of the proposed sampling model since it is appropriate for use in conjunction with designs incorporating experimental methods which are available for educational research and for designs concerned with the evaluation of the influences of educational programs.

A Reconsideration of the Cohort Variable

We have suggested previously that the definition of the cohort variable need not be restricted to date of birth as Schaie (1965) has assumed. Such a definition is most appropriate, perhaps for studies concerned with the behavior and development of infants (e.g., Weatherford & Cohen, 1973). However, even in these instances, the definition can be called into question. As an example, Fantz, Fagan and Miranda (1975) have suggested that date of conception, rather than date of birth, is a more appropriate index by which to identify the "origin" of life. Similarly, genetic influences on behavior assuredly profit from a definition of cohort based on family lineage. Baltes and Rienert (1969) and Buss (in press), and others have also provided compelling discussions which question the interpretations of "cohort" effects drawn from studies adopting Schaie's definition. Like age, the cohort variable can take many forms having a biological, sociological, or psychological basis. For example, cohort can be defined by social or environmental factors which are shared by a specific segment of society at the same time (e.g., entrance into school, graduation, etc.) or by a society as a whole (e.g., war, depression). Matters are made even more complex when it is considered that many of these events are correlated with CA, time of measurement, and date of birth. For example, the social state of marriage is correlated with age in the general population but nevertheless may have pronounced behavioral correlates which exist either independently or in interaction with age.

LONG-TERM DEVELOPMENTAL RESEARCH

Birren (1959) noted the absence of developmental scales which reflect biological, psychological, or sociological "age" over the long term, and Wohlwill (1973) has recently reiterated this conclusion. For this reason chronological age continues to serve as the predominant criterion for subject selection and matching in developmental research. It is important to note that the reasons for using chronological age vary widely across different researchers and different studies. For example, CA may be used bacause our society is "age graded," because CA correlates with biological, or psychological development, etc. Nevertheless, such relationships are not necessarily stable over the long term (e.g., Neugarten & Moore, 1968).

A second point is that very little developmental research is concerned with behavior change over a large segment of the life span. Impediments to life-span research have included the artificial segmentation of the life-span as well as the failure of developmental theories to encompass a whole-life perspective.

In addition, it has been noted here that developmental researchers rarely attend to the point of origin as a nominal property of a CA-based scale. Rather, development (i.e., behavior change over time) is examined in relation to the developmentally "youngest" sample included in the investigation. The suggestion here is that developmental change is most properly assessed in relation to a sample selected and defined in terms of process-defined criteria directly related to the theory or hypotheses central to the investigation. Thus, the segment of the life span which is sampled in a developmental study may be restricted to the period over which the process is assumed to influence behavior. Another implication is that the construction and use of developmental scales based on process-related criteria need not encompass the life-span unless the process itself is assumed to be of central importance across this period. Long-term developmental changes in behavior may not be properly represented using a single developmental scale. More important for present purposes, however, is that shorter-term changes may be efficiently described through the selection of a scale defined by a functional point of origin and a metric of time in the manner illustrated in Tables 2-5.

SUMMARY

Chapters 2 and 3 have highlighted the methodological complexities involved in the conduct of research concerned with studying B = f(T) phenomena. The attempts to resolve the complexities through the use of sequential sampling strategies such as those provided by Schaie (1965) and Baltes (1968) must be viewed as very significant advancements. However, it has been shown that the use of a sequential design (as a replacement for the longitudinal, cross-sectional, or time-lag design) is no panacea unless the hypothesis guiding the study of the B = f(T) phenomena of interest are firmly grounded in theory. Furthermore, the theory guiding the investigation should specify the underlying scale along which the B = f(T) phenomena change and the major factors (e.g., age, time-of-measurement, or cohort) influencing behavior and performance for the time period, social context, and population being studied. The theory should also provide strong direction to the researcher in selecting the times of testing and the ages of children from which to collect data. Finally, the theory must specify the relation between the factors of age, time-of-measurement, and cohort. It is only when this is accomplished that a sampling model conforming to Schaie's general developmental model or the use of one of the Schaie (1965) and Baltes (1968) can be selected as the optimal sampling strategy for the behaviors being studied. The controversy between Schaie (1965) and Baltes (1968) as to whether Schaie's model conforms to a trifactor or bifactor model is a case in point which can only be settled in the context of a theory which speaks directly to these issues and those discussed in this section.

REFERENCES

Anderson, S. B., & Maier, M. H. 34,000 pupils and how they grow. Journal of Teacher Education, 1963.

Austin, G. R., Rogers, B. G., & Walbesser, H. H., Jr. The effectiveness of summer compensatory education: A review of the research. Review of Educational Research, 1972, 42, 171-182.

Baltes, P. B. Longitudinal and cross-sectional sequences in the study of age and generation effects. Human Development, 1968, 11, 145-171.

Baltes, P. B., & Goulet, L. R. Exploration of developmental variables by manipulation and simulation of age differences in behavior. Human Development, 1971, 14, 149-170.

Baltes, P. B., & Reinert, G. Cohort effects in cognitive development of children as revealed by cross-sectional sequences. Developmental Psychology, 1969, 1, 169-177.

Birren, J. E. The psychology of aging. New York: Prentice-Hall, 1964.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963, 171-246.

Coleman, J. S., et al. Equality of educational opportunity, Catalog No. FS 5.238:38000. Educational Testing Service Annual Report. Princeton, N. J.: Educational Testing Service, 1964.

Fantz, R. L., Fagan, J. F., & Miranda, S. B. Early visual selectivity as a function of pattern variables, previous exposure, age from birth and conception, and expected cognitive deficit. In L. Cohen and P. Salatatek (Eds.), Infant Perception. Vol 1. New York: Academic Press, 1975.

Gelman, R. Conservation acquisition: A problem of learning to attend to relevant attributes. Journal of Experimental Child Psychology, 1969, 7, 167-187.

Goulet, L. R. Verbal learning in children: Implications for developmental research. Psychological Bulletin, 1968, 69, 359-376.

Goulet, L. R. Training, transfer, and the development of complex behavior. Human Development, 1970, 13 (4), 213-240.

Goulet, L. R., & Goodwin, K. S. Development and choice behavior in probabilistic and problem-solving tasks. In H. W. Reese & L. P. Lipsitt (Eds.), Advances in child development and behavior, V, New York: Academic Press, 1970, 213-254.

Goulet, L. R., Williams, K. G., Bozinou, E., & Hexner, P. Z. Longitudinal and time-lag differences in rule utilization schooling and age-related effects. Unpublished manuscript, 1973.

Goulet, L. R., Williams, K. G., & Hay, C. M. Age- and schooling-related changes in memory performance. Unpublished manuscript, 1973.

Goulet, L. R., Williams, K. G., & Hay, C. M. Longitudinal changes in intellectual functioning in pre-school children: Schooling and age-related effects. Journal of Educational Psychology, 1974, 66, 657-662.

Hilton, T. L., & Meyers, A. E. Personal background, experience, and school achievement: An investigation of the contribution of questionnaire data to academic prediction. Journal of Educational Measurement, 1967, 4, 69-80.

Hilton, T. L., & Patrick, C. Cross-sectional versus longitudinal data: An empirical comparison of mean differences in academic growth. Journal of Educational Measurement, 1970, 7, 15-24.

Kessen, W. Research design in the study of developmental problems. In P. H. Mussen (Ed.), Handbook of research methods in child development. New York: Wiley, 1960, 36-70.

Neugarten, B. L., Moore, J. W. & Lowe, J. C. Age norms, age constraints, and adult socialization. American Journal of Sociology, 1965, 70, 710-717.

Schaie, K. W. A general model for the study of developmental problems. Psychological Bulletin, 1965, 64, 92-107.

Schaie, K. W. Limitations on the generalizability of growth curves of intelligence: A reanalysis of some data from the Harvard Growth Study. Human Development, 1972, 15, 141-152.

Sigel, I. E., & Hooper, F. H. Logical thinking in children: Research based on Piaget's theory. New York: Holt, Rinehart and Winston, 1968.

Weatherford, M. J. & Cohen, L. B. Developmental changes in infant visual preferences for novelty and familiarity. Child Development, 1973, 44, 416-424.

Weir, M. W. Developmental changes in problem-solving strategies. Psychological Review, 1964, 71, 473-490.

Wohlwill, J. F. The age variable in psychological research. Psychological Review, 1970, 77, 49-64.

Wohlwill, J. The study of behavioral development. New York: Academic Press, 1973.

Wood, P. A., & Goulet, L. R. Age and school experience as factors related to visual perception. American Educational Research Journal, 1973a.

Wood, P. A., & Goulet, L. R. Longitudinal and grade-related differences in visual-perceptual performance. Unpublished manuscript, 1973.

CHAPTER 4

THE DETERMINATION OF THE SIGNIFICANCE OF CHANGE BETWEEN PRE AND POSTTESTING PERIODS

The measurement of change has been a favorite topic of psychometricians for years. It is a topic with considerable problems many of which are best avoided by following the advice of Cronbach and Furby (1970) to "...investigators who ask questions regarding gain scores..." that they "...frame their questions in other ways" (p. 80).

In many situations, gain scores appear to be the natural measure to be obtained. In some instances, however, the formulation of the questions in terms of gains introduces unnecessary problems. In other instances the gain formulation gives the illusion that certain types of inferences can be made when in fact they are not justified. In the latter case, the gain formulation conceals limitations that are inherent in the data.

In this chapter some of the major issues that arise in the measurement of change are reviewed and, where possible, alternative approaches are discussed. The measurement of individual differences is considered first. This is followed by a discussion of some of the concerns involved in inferring treatment effects from group differences. The chapter is then concluded with a section on accountability systems based on student achievement.

INDIVIDUAL DIFFERENCES

Some of the best known problems in the measurement of change arise in situations where there is an interest in measuring individual differences. It may be desired to identify individuals who gain unusually large (or small) amounts so that these individuals may be given special treatment. In the case of some performance contracts, individual gain scores have been used as the basis of determining payment to contractors. In other situations there may be an interest in identifying the correlates of change. While not involving individual change scores, as such, correlational uses of change scores are also considered under the heading of individual differences.

Difference Scores

The most natural measure of change from one point in time to another is the simple difference score. The dieter quite naturally is interested in the difference between his pre diet weight and his post diet weight. It is somewhat ironic that this simple procedure results in a score with several major defects.

Negative correlation with pretest (e.g., Bereiter, 1963; Thorndike, 1966): A major disadvantage of the simple difference score is that it typically has a negative correlation with the pretest. The correlation

of a pre measure, X, with the difference between a post measure, Y, and that pre measure is

$$\rho_{XD} = \frac{\rho_{xy}\,\sigma_y - \sigma_x}{\sqrt{\sigma_x^2 + \sigma_x^2 - 2\rho_{xy}\,\sigma_x\,\sigma_y}} \qquad (4.1)$$

where $D = Y - X$, σ_x and σ_y are the standard deviations of X and Y respectively, and ρ_{xy} is the correlation between X and Y. It is clear from an inspection of the numerator in equation 4.1 that the correlation between D and X will be negative unless $\rho_{xy}\,\sigma_y$ is greater than σ_x. Typically, $\rho_{xy}\,\sigma_y$ will be smaller than σ_x because the correlation between X and Y must be less than one and the standard deviations of the pre and post measures are often of relatively similar magnitude. Although there is a tendency for the correlation to be negative it is, of course, possible for the correlation to be positive but only if the standard deviation of the post measure, Y, is larger than that of the pre measure, X, and generally substantially so. It should also be noted that since the two terms in the numerator of equation 4.1 are of opposite sign, the magnitude of the correlation will usually be small in absolute value.

An implication of the negative correlation between D and X, is that large positive D's are more likely to be observed for persons with low X scores whereas persons with high X scores would have large positive D's only rarely. Thus, if individuals with high D scores are to be selected, there will be an overrepresentation of people with low X scores as an artifact due to the negative correlation between D and X.

Low Reliability (e.g., Lord, 1963): Given the standard assumptions of classical test theory, the reliability of a difference score is

$$\rho_{DD'} = \frac{\rho_{xx'}\,\sigma_x^2 + \rho_{yy'}\,\sigma_y^2 - 2\rho_{xy}\,\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 - 2\rho_{xy}\sigma_x\sigma_y} \qquad (4.2)$$

where $\rho_{xx'}$ and σ_x^2 are the pretest reliability and variance, $\rho_{yy'}$ and σ_y^2 are the posttest reliability and variance, and ρ_{xy} is the correlation between pre and posttests. Consider the special case of 4.2 where

$$\sigma_x = \sigma_y \text{ and } \rho_{xx'} = \rho_{yy'} = \rho$$

then $\rho_{DD'}$ can be written

$$\rho_{DD'} = \frac{\rho - \rho_{xy}}{1 - \rho_{xy}}. \qquad (4.3)$$

Although 4.3 applies only for a specialized situation it may be instructive to consider values of $\rho_{DD'}$ for selected values of ρ and ρ_{xy}. This is done in Table 4-1 and as can be seen there, the value of $\rho_{DD'}$ is discouragingly low when ρ_{xy} is at all large.

Of course, one way to obtain more reliable difference scores is to have a low correlation between pre and post scores. Under such circumstances, however, it is questionable that the pre and post measures are getting at the same construct which would seem to be a prerequisite for the difference score to be interpreted as an index of growth.

An implication of the low reliability of difference scores is that it is quite risky to make any important decisions about individuals on the basis of gains from pre to post testing periods. A practical situation where the low reliability of a difference score causes problems is that of performance contracting. Even without any real change it is possible to find substantial numbers of individuals with large difference scores due simply to the low reliability of these scores. Stake (1971) has illustrated this problem. He concluded that "...owing to unreliability, gain scores can appear to reflect learning that actually does not occur" (p. 587).

Lack of Common Trait and Scale (e.g., Bereiter, 1963; also Chapters 5 and 7 of this report): It would hardly be sensible to estimate a person's gain in weight by subtracting the number of pounds he weighed at time 1 from the number of ounces he weighed at time 2. To make sense the same scale units must be used at both points in time. Similarly, it would make no sense to subtract a pre measure of height from a post measure of weight to get an estimate of weight gain. It is necessary to measure weight at both points in time.

The need for a common scale and trait at pre and posttesting periods which is so obvious with the above physical examples is sometimes less obvious, but no less essential, in an educational context. For example, if arithmetic test A was used as the pre measure and arithmetic test B as the post measure it might be forgotten that the units of the two tests are unequal. Even more likely, it might be forgotten that test A consists primarily of addition problems while test B consists largely of subtraction problems. Under such conditions the difference scores would not necessarily be measuring gains along the dimension measured by test A any more than the difference scores in the two examples involving weight measure weight gain. Even where the same test (or parallel forms) is used as the pre and post measures it is sometimes the case that different constructs are measured at the two points in time. For example, an item which measures problem solving skill at one point in time may measure memory at a later point in time.

Residual Scores

Problems inherent to difference scores have led a number of people to seek altenatives. One of these is the residual score which is largely motivated by the desire for a score that has a zero correlation with the

Table 4-1

Difference Score Reliability as a Function of the Reliability of the Parts and Their Intercorrelation*

Correlation of Pre and Post Score	Reliability of Pre and of Post Scores (assumed to be equal)		
	.7	.8	.9
.5	.40	.60	.80
.6	.25	.50	.75
.7	.00	.33	.67
.8	—	.00	.50
.9	—	—	.00

*Assuming $\rho_{xx'} = \rho_{yy'}$ and $\sigma_x = \sigma_y$.

pretest (DuBois, 1957; Manning & DuBois, 1962). As noted by Cronbach and Furby (1970) "One cannot argue that the residualized score is a 'corrected' measure of gain..." rather it "...is primarily a way of singling out individuals who have gained more (or less) than expected" (p. 74).

A residual score, R, is obtained by subtracting the predicted posttest score, Y, from the corresponding observed posttest score, Y. The predicted posttest score is obtained from the linear regression of Y on the pretest, X. The zero correlation between X and R follows immediately from the way in which R is derived and is seen as a major advantage over difference scores because residuals do not give an advantage to persons with certain values of the pretest scores whereas difference scores do.

While solving the problem caused by the correlation between difference and pretest scores residuals, like difference scores tend to be unreliable. As indicated by O'Connor (1972) the reliability of a residual score can be written as

$$\rho_{RR'} = \frac{\rho_{yy'} - \rho_{xy}^2(2 - \rho_{xx'})}{1 - \rho_{xy}^2}.$$

Values of the reliability of residual scores are reported in Table 4-2 for selected values of ρ and ρ_{xy} under the assumption that $\rho_{xx'} = \rho_{yy'} = \rho$. The values of ρ and ρ_{xy} used in Table 4-2 are the same as those used in Table 4-1.

Although the residual score reliabilities shown in Table 4-2 are somewhat better than the corresponding difference score reliabilities shown in Table 4-1, they are still disappointedly small whenever the correlation of pre and post scores is large. Furthermore, residuals are usually of most interest in situations where the pre-post correlation is large relative to the reliabilities of the parts. Thus, the same cautions due to unreliability of difference scores also apply to residual scores.

Estimated True Change

Another alternative to the raw difference score approach is to estimate "true" change. In other words, the change that would be obtained if there were no errors of measurement is estimated. The true change is presumably the quantity of real interest whenever an attempt is made to measure change.

In the case of a single measure there is a perfect correlation between the estimated true score for that measure and the observed score. Hence, for most purposes the observed score serves just as well as the estimated true score. Whenever two or more measures are available, however, the estimated true score based on all available information will ordinarily have a less than perfect correlation with the observed score of the measure. In the case of a difference score both the pretest and the posttest, and if available, other scores as well provide information about the true difference score and the resulting estimated true score may result in noticeably

Table 4-2

Residual Score Reliability as a Function of the Reliability of the Parts and Their Intercorrelations*

Correlation of Pre and Post-Scores	Reliability of Pre and of Post Scores (assumed to be equal)		
	.7	.8	.9
.5	.50	.67	.83
.6	.36	.58	.79
.7	.12	.42	.71
.8	—	.09	.54
.9	—	—	.05

*Assuming $\rho_{xx'} = \rho_{yy'}$.

different ranking of individual than would be obtained from raw difference scores.

Regression Estimates (Lord, 1956, 1958, 1963; McNemar, 1958; Cronbach and Furby, 1970; Marks & Martin, 1973): Given estimates of the reliabilities of the pretest and of the posttest as well as their variances and their covariance, it is possible to obtain estimates of true gain using multiple regression procedures. The basic formulas may be found in Lord (1963, p. 28). Cronbach and Furby (1970) extend these formulas by distinguishing between linked measures (i.e., ones with correlated errors) and independent measures. They also consider the possibility of using other available measures as predictors.

As Lord (1963) has shown with an empirical example, persons with the largest estimated true difference scores are not necessarily those with the largest observed difference scores. In particular, persons with relatively large pretest scores are more apt to be among those with "large" gains when estimated true difference scores (Lord, 1963, equation 3) are used than when raw difference scores are used. Thus, the estimated true difference scores obviate the objection that difference scores tend to favor persons with low pretest scores.

As noted by Cronbach and Furby (1970), it is not necessary to limit the estimation to the measures involved in the difference score. Any measures that are available may be used along with the pre and the post measures to estimate the true difference score. As shown by Tatsuoka (1975), the additional measures will improve the prediction of the true difference if they are correlated with the errors of measurement on X and/or Y. In practice, the addition of more predictor variables would probably improve the accuracy of the estimate relatively little unless the pre and post measures were of low reliability.

The reliability of estimated true change is equal to the squared multiple correlation of true change with the predictor variables, i.e., with X, Y and possibly other measures. It will always be as large or larger than the reliability of a simple difference score (Tatsuoka, 1975). When $\rho_{xx'} = \rho_{yy'}$ and $\sigma_x = \sigma_y$ the reliability of the estimated true difference scores equals that of the raw difference scores (see equation 4-3). If the pre and posttest reliabilities and/or the pre and posttest variances are unequal then the reliability of the estimated true difference scores will exceed that of raw difference scores but typically only slightly. For example, if $\rho_{xx'} = .85$, $\rho_{yy'} = .90$, $\sigma_x = 1.5$, $\sigma_y = 1.7$, and $\rho_{xy} = .7$ then the reliability of the raw difference score computed from (2) is .600 which can be compared to the reliability of the estimated true difference score of .613.

Linked vs. Independent Observations (Cronbach & Furby, 1970; Werts, Jöreskog & Linn, 1972): All of the preceding discussion depends on the usual assumptions of classical test theory. In particular, it is implicitly assumed that the pretest errors of measurement are uncorrelated with the

posttest errors of measurement. Where the same instrument is used to obtain both pre and post measures the assumption of uncorrelated errors of measurement may be especially dubious. Thus, it is desirable to use estimation procedures that allow for the possibility of correlated errors of measurement on the pre and posttest. To do so, however, requires the availability of more information in the form of multiple measures than is often available in practical settings.

Cronbach and Furby (1970) have formalized the distinction between linked and independent observations. They distinguish two types of error components. For linked observations (e.g., the same form of a test used as both the pre and the posttest) one type of error component would be assumed to have a nonzero correlation. On the other hand, independent observations would be assumed to have both types of error components uncorrelated. The distinction between linked and independent observations leads to different formulas for estimating the reliabilities of difference scores and true change. Basically the formulas require that a distinction be made between the correlation of X and Y where X and Y are linked and where X and Y are independent observations. Furthermore, separate estimates of the linked ρ_{xy} and the independent observations ρ_{xy} are required.

Correlates of Change

Frequently the focus in measuring change is not on the individual difference scores but on their correlates. The interest is in finding variables that predict the amount of change. Measures of change may sometimes be computed for individuals as a means to the end of correlating these measures with other variables. Frequently, however, the change measures need not actually be computed to obtain the desired correlations of these measures with other variables.

The alternative approaches to measuring change result in different correlations of these measures with other variables. The different estimates have different theoretical and practical implications.

Spurious Correlations (Lord, 1963). Earlier the tendency for a difference score to have a negative correlation with the pretest was noted. More generally, the correlation of a raw difference score with another variable that is partially a function of the pretest or posttest is usually considered spurious (Lord, 1963, p. 33). The spuriousness is the result of the same errors of measurement occurring in the difference score and in the variable with which it is correlated. In the case of the correlation of D = Y - X with X the same errors of measurement that are positively weighted for X are negatively weighted for D and the result is usually a spurious negative correlation.

Attenuation (Lord, 1963): Unreliability has the effect of attenuating correlations. This is true of all fallible measures but becomes of major importance when the reliability of a variable is quite low as is typically the case for measures of change. The practical implication of the large degree of attenuation that is typically encountered with difference scores is that correlations involving a difference score will tend to be quite

low which is rather discouraging for someone who is interested in finding correlates of change.

Part and Partial Correlations: If residual scores rather than difference scores are used in correlational studies, the result is the same as a part correlation. That is, the pretest score, X, is partialled out of the posttest score, Y, and the residual is correlated with a third variable, W. Note that X is not partialled out of W but only out of Y. The result is called a part correlation. Thus, X is held constant statistically with respect to Y but not with respect to W.

A more familiar correlational approach is to partial X out of both Y and W. The result is called a partial correlation and has a somewhat simpler interpretation than the part correlation since X is held constant statistically with respect to both Y and W instead for just one of them as in the case of part correlation. If X, Y and W have a multivariate normal distribution, then the partial correlation of Y and W with X partialled out is simply equal to the correlation between W and Y for any fixed value of X. This would often seem to be a coefficient of interest where the focus is on correlates of change from pre to posttesting periods. As previously noted, however, residual scores cannot be considered as better measures of change. They merely represent that part of a score that is not linearly predictable from the variable that is partialled out. Nonetheless, the partial correlation provides a means of identifying variables that can predict posttest scores of individuals with equal pretest scores.

The problem of unreliability that runs throughout the measurement of change is also a major concern with partial correlation. The direction of the effect of unreliability on a simple correlation is known in advance. Unfortunately, this is not true of partial correlations (Lord, 1963, p. 36). In the case of partial correlations, the effect of errors of measurement may be to change the sign of a partial correlation. As shown by Linn and Werts (1973) it is possible for errors of measurement to result in a partial correlation of zero where the partial correlation among the error free measures is non-zero. For these reasons, it is particularly important to make corrections for attenuation when using partial correlations.

Partial Regression Weights: Within the context of a linear model, the relationship of a variable, W, with change might be evaluated in terms of the regression of the change on W and the pretest. Werts and Linn (1970) have shown that the resulting partial regression weights can be readily obtained from the partial regression coefficients in the regression of the posttest on W and X. Hence, there is no need to actually use difference scores. This is true with or without corrections for unreliability of the measures.

Recommendations (Individual Differences):

One of the most common uses of change measures is as criteria in correlational studies. The goal of such studies is the identification of variables that predict who will gain the most in a particular situation.

Cronbach and Furby (1970) argue that it is preferable to phrase such questions in terms of partial correlations rather than correlations involving difference scores or in terms of part correlations. We concur with this recommendation. Regardless of the way in which such questions are phrased, however, it is important to take the unreliability of the measures into account.

In the case of partial correlations, taking the unreliability into account "...poses somewhat of a dilemma, since, first, it is often hard to obtain the particular kind of reliability coefficients that are required for making the appropriate correction, and, further, the partial corrected for attenuation may be seriously effected by sampling errors. These obstacles can hardly justify the use of an uncorrected coefficient that may have the wrong sign, however, (Lord, 1963; 36)."

Two other possible cases of change measures relating to individual differences that are discussed by Cronbach and Furby (1970) are the identification of individuals with unusually large (or small) gains and the use of change measures as theoretical constructs. In neither case are change scores needed. In the former case the regression approach outlined by Cronbach and Furby is preferred. In the latter case, linear combinations other than simple difference scores, with the arbitrary weights of plus and minus one, should be allowed (Cronbach and Furby, 1970).

GROUP DIFFERENCES (INFERRING TREATMENT EFFECTS)

Questions about the effects of experimental treatments or of variables involved in observational studies are frequently phrased in terms of gains. For example, does treatment A result in a larger gain than treatment B? Do students in integrated schools gain more than students in segregated schools? Do students in "open" classrooms gain more than those in "traditional" classrooms? Although these questions seem intuitively reasonable it does not follow that the best approach to trying to answer them will involve the use of measures of change as dependent variables. Indeed, "...There appears to be no need to use measures of change as dependent variables and no virtue in using them (Cronbach and Furby, 1970, p. 78).

An important distinction among investigations aimed at inferring treatment effects must be made between studies that have random assignment and those that don't. For studies with random assignment a pretest serves primarily as a means of increasing statistical power. Where treatment groups are not formed by random assignment it is often hoped that the pretest will provide a means of allowing for preexisting differences.

Random Assignment

When treatment groups are formed by randomly assigning individuals (or more generally units) to treatment conditions, the posttest alone is perfectly suitable as a dependent variable. A test of the null hypothesis of equal posttest means for the treatment groups is appropriate for evaluating treatment effects. If pretest measures are available in this context their potential usefullness is best evaluated in terms of the effect of each use on the power of the statistical test.

A pretest may increase the precision of an experiment. The extent to which experimental precision is improved depends on the way in which the pretest information is used as well as on the nature and magnitude of the pretest-posttest relationship. Difference scores are one possibility but not the only one. Feldt (1958) compared three potential uses of concomitant variables: (1) blocking, (2) analysis of variance on difference scores, and (3) analysis of covariance. He clearly shows that among these three approaches that the difference score approach has the least precision. Thus, on the basis of precision the choice would ordinarily be between blocking and the analysis of covariance with the analysis of covariance being the most precise where the correlation between the pre and posttest is greater than .6 (Feldt, 1958).

In pretest-posttest designs the correlation between the pretest and the posttest is frequently .7 or higher. Thus, the analysis of covariance would seem to be an attractive approach to the analysis of such data. Before this technique is wholeheartedly accepted, however, several limitations of the technique need to be considered. As Elashoff (1969) has argued, the analysis of covariance is a "delicate instrument". Elashoff notes that the analysis of covariance involves a number of relatively strong assumptions and violations of some of these assumptions may invalidate the technique. Where the assumptions of linearity or of homogeneity of regression seem questionable it may be preferable to use the pretest as a blocking variable rather than a covariate. In any event, however, there seems to be little justification for using difference scores.

Another assumption of the analysis of covariance is that the covariate is measured without error. Violations of this assumption are most troublesome in situations where groups are not formed by random assignment and will be considered again in that context. Even with random assignment errors of measurement limit the value of traditional analysis of covariance. But, techniques are available for allowing for errors of measurement in the covariate (Lord, 1960; Porter, 1967).

Preformed Groups

Random assignment is seemingly impossible in many situations where answers to questions about treatment effects are sought. Children cannot ordinarily be randomly assigned to schools or to major programs such as Head Start. Even if such random assignment were administratively feasible, it might not be desirable on grounds other than the desire for a clean experimental design. Without random assignment it is, of course, possible that differences that may be observed in the posttest score are the result of preexisting group differences rather than treatment effects. What is desired is a means of allowing for preexisting group differences. It is the hope of achieving this goal that often leads to the use of difference scores or the analysis of covariance.

Lord (1967, 1968) has provided a compelling analysis of the use of difference scores or the analysis of covariance to infer treatment effects from studies involving preformed groups. He has clearly shown that the

two approaches can lead to contradictory results. The basic problem is one of making the proper adjustment for any preexisting differences. Unfortunately, there is no way of knowing which of these or any other techniques provide the proper adjustments. According to Lord "...there simply is no logical or statistical procedure that can be counted on to make proper allowances for uncontrolled preexisting differences between groups (1967, p. 305)."

This discouraging conclusion is also reached by Meehl (1970) and by Cronbach and Furby (1970) among others. Without assurance of proper adjustments for preexisting differences, there is necessarily a concern about the possibility that treatment effects, however obtained, may be subject to major sources of bias. In order to evaluate the bias in various nonexperimental research situations it is important to have a clear understanding of what is meant by a treatment effect. Rubin (1972) has provided a definition which is useful from a formal point of view as well as being consistent with intuitive notions of a "causal effect." His basic definition of an effect is specific to each unit (e.g., individual student, classroom, school) under consideration, to a particular time interval (t_1 to t_2) and to a particular pair of treatments (e.g., experimental and control). The effect of the experimental versus the control treatment on a dependent variable, X, is the difference between the score on X that would have been obtained by the unit at t_2 if the experimental treatment had been introduced at t_1 and the score on X that would have been obtained by the unit at t_2 if the control treatment had been introduced at t_1.

In practice it is impossible to measure the effect defined above for any unit because only one treatment can be introduced at t_1 and it is impossible to return to that time to introduce the other treatment. Nor is it possible to measure the average effect of all units for the same reason. Nonetheless this formulation is useful because under random assignment of units to treatments, the expected value of the difference in mean scores on X is equal to the average difference that would be observed if all units could be observed under both treatment conditions during the same time interval. Thus, the sense in which the randomized experiment provides an unbiased estimate of the treatment effect is clear. Furthermore, a framework is provided for considering factors in nonexperimental designs that result in biased estimates. In this way it may sometimes be possible to specify conditions under wich estimated treatment effects may be biased in one direction or the other or to clearly specify the a priori assumptions that would have to be satisfied for the estimate to be unbiased.

One of the many potential sources of bias in estimated treatment effects from the analysis of covariance is due to errors of measurement in the pretest (Porter, 1967; Werts and Linn, 1971). The effect of unreliability in the covariate is a reduction in the slope of the regression of the dependent variable on the covariate. Where there are preexisting

differences in the group means on the covariate the reduction in slope leads to bias in the estimated magnitude of the treatment effects. The direction of the bias due to unreliability of the covariate can be determined and if adequate estimates of the cavariate reliability can be obtained, the procedures outlined by Porter can be used.

Single Group Designs

For a single group such as a school or classroom there may be an interest in the amount of change that occurs during a given time interval. Once again there is no real virtue to difference scores (Cronbach and Furby, 1970). A simple t-test for dependent samples will provide a test of the null hypothesis that the mean pretest score equals the mean posttest differences.

While such differences may be due to the school experience they might also be due to a host of non-school experiences that students have during the interval between the pre and posttests. An observed difference may be attributable to variables associated with increased chronological age which have nothing to do with school effects per se. It would be desirable to separate differences in test scores that are associated with chronological age. Goulet (in press) has proposed an approach that is specifically designed for this purpose.

Goulet (in press) suggested a sampling procedure that would provide for independent estimates of effects associated with chronological age and those associated with amount of schooling as well as their interaction. His design would require that nonoverlapping random samples of students be tested at different points in the school year. The students' scores would then be categorized according to chronological age and time of testing. A simple design involving four different samples of children is shown below.

Age at Testing Date	Time of testing Sept.	Jan.
7-3	A	B
7-7	C	D

The means based on subsamples A, B, C and D above provide the basis for estimating effects associated with schooling that are independent of affects associated with age. As indicated by Goulet, (in press) the desired estimate is simply

$$\frac{\bar{X}_B + \bar{X}_D - \bar{X}_A - \bar{X}_C}{2}$$

where the $\bar{X}$'s refer to the subsample means. Goulet's suggested approach does not guarantee that the estimated effect is due to school. It still might, for example, be the result of factors outside the school experience

which covary with that experience. It demonstrates, however, an approach for separating two major sources of competing hypotheses about clusters of variables that might influence pupil performance. By holding constant sources of variance associated with age, the estimates of "school effects" are much more compelling than when the estimates involve a combination of school associated and age associated effects. A more complete discussion of sampling designs such as the above is provided in Chapter 2.

ACCOUNTABILITY SYSTEMS BASED ON STUDENT ACHIEVEMENT

There may be fairly general agreement with the conclusion stated by Lord (1967) that there is generally no way of knowing what adjustments should be made to allow for preexisting group differences. Nonetheless many practical decisions must be made without the aid of randomized experiments. These decisions must be made on some basis. Even with all of the pitfalls that are encountered in trying to interpret information that can be gleaned from data collected for preexisting groups, it still often seems to be the best alternative.

Responses to pressures to be accountable have taken many forms. Educational accountability has many meanings and as Glass (1972) has indicated not all of the uses of the term require the measurement of student performance. One of the more common interpretations, however, is that educators should be accountable for what students learn. For this interpretation of accountability the results of standardized achievement tests would seem a natural source of information not only for assessing current status but for evaluating progress. Unfortunately, there is great potential for misuse of standardized test results for purposes of educational accountability.

Norms as Standards

Knowing only a student's raw score on a test would provide essentially no information. To derive meaning the content of the items must be known in some detail. If the content is described in sufficient detail then a statement that a student got 20 of 40 items correct would begin to take on some meaning but would still not be a sufficient basis for answering a parent's question about whether that was good. There are two major approaches that are commonly taken to answering this question: criterion referenced and norm referenced. The more common of these is the norm referenced approach which simply provides a comparison of the student's performance to some specified group. The norms may take the form of percentile ranks, grade equivalents or some other type of scaled score but basically the norms provide a means of interpreting a students' performance relative to that of other students.

Grade Equivalent Scores: A problem with the use of norms is that the norm is sometimes confused with the standard or ideal. It is obvious that not all children can be above the 50th percentile. It should be just as obvious that not all schools can be above the 50th percentile of school mean norms. When grade equivalents are used it is still the case that not all children (or schools) can be above grade level but this may be less obvious with grade equivalent scores than with some other types of scales.

The grade equivalent score suffers from a number of defects (see for example Angoff, 1971). Most of these defects stem from the surplus meaning that is attached to the label. Because of these defects in the grade equivalent the latest version of the Standards for Educational and Psychological Tests recommends that they be discontinued or their use discouraged (APA, 1974).

Change on Achievement Test Scales. Regardless of the nature of the scale that is used, scores at a single point in time could hardly be expected to provide information about the effectiveness of a school. The notion that educators should be accountable for student learning has implicit in it the notion of change. True, a measurement at a single point in time may provide information about strengths or weaknesses but it cannot be expected to indicate by itself the amount of progress that was made in any given interval of time. To do this something must be known about past as well as present performance. The desire to know something about progress brings us back to our concern about change from pre to posttesting periods.

Probably the most widely used scale for purposes of evaluating pupil growth is the grade equivalent scale (see for example, Wargo, et al., 1972). The deceptive simplicity of grade equivalents makes them appear particularly useful for the purpose of measuring growth. Lindquist and Hieronomous, for example, say that "Grade equivalent scores are best suited for measuring growth from year to year (1964, p. 13)."

Although Lindquist and Hieronomous go on to discuss limitations of grade equivalent scores, these limitations are often overlooked. One of the potentially misleading characteristics of grade equivalents is that they seem to provide a standard of "normal" growth. If educational accountability is interpreted to mean that someone should be responsible for the progress or lack of progress displayed by students, then some notion of satisfactory progress is needed. To many people, the grade equivalent seems to provide the standard. That is, the gain of one grade equivalent in a year's time becomes the standard to be expected. Unfortunately, however, "...a year's progress in a year's time means different things to a teacher whose class begins the year near or above grade level and a teacher whose class begins two or three years below grade level (Rosenshine and McGaw, 1972, p. 640)."

Some of the problems encountered in trying to interpret gains on standardized achievement scales may be illustrated by the following example results from a school system. An attempt was made to look at the gain in achievement test performance for students in three broad categories of ability as measured by IQ test scores. Standardized achievement test data were obtained for students in grades 3 and 6. Results were also obtained for these same students the following year when they were in grades 4 and 7. Grade scores or grade equivalents were then reported in reading and in arithmetic at each point in time and gain scores were computed over the one-year interval. The mean scores and mean gains were reported separately by school and for students with IQ's of 114 or above, those with IQ scores of 98 to 113, and those with IQ's of 97 or less. This was done for each school and for the school system as a whole.

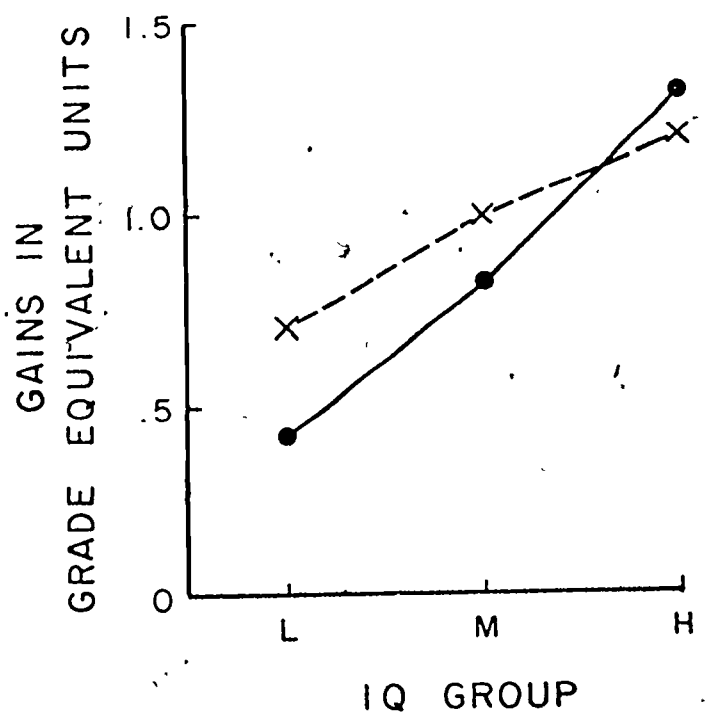
For the school system as a whole, the gains for each of the IQ levels (L, M and H) are plotted in Figure 1 for reading and for arithmetic. Section (a) of Figure 4-1 shows the results for 3rd to 4th grade gains on the Metropolitan Achievement Test (Harcourt Brace Jovanovich, 1970). The gains observed for 6th to 7th grade are based on the Educational Development Series (Scholastic Test Service, 1969; 1971).

From section (a) of Figure 4-1 it can be seen that from grades 3 to 4 the largest gains in both reading and arithmetic were made by the high IQ group and the smallest gains by the low IQ group. As would be expected, the high ability students had a higher mean test score on the pretest than the low ability students. At the time of the second test the gap between the two extreme groups of students had widened. In reading, the gap between the two groups was 1.5 GE units at grade 3 and 2.4 GE units at grade 4. The result is quite consistent with the expectation that "the rich get richer and the poor get poorer." It is also consistent with the results that have been reported indicating that, as measured by standardized tests, the gap in achievement between high and low SES or between minority and majority groups tends to increase with grade level.

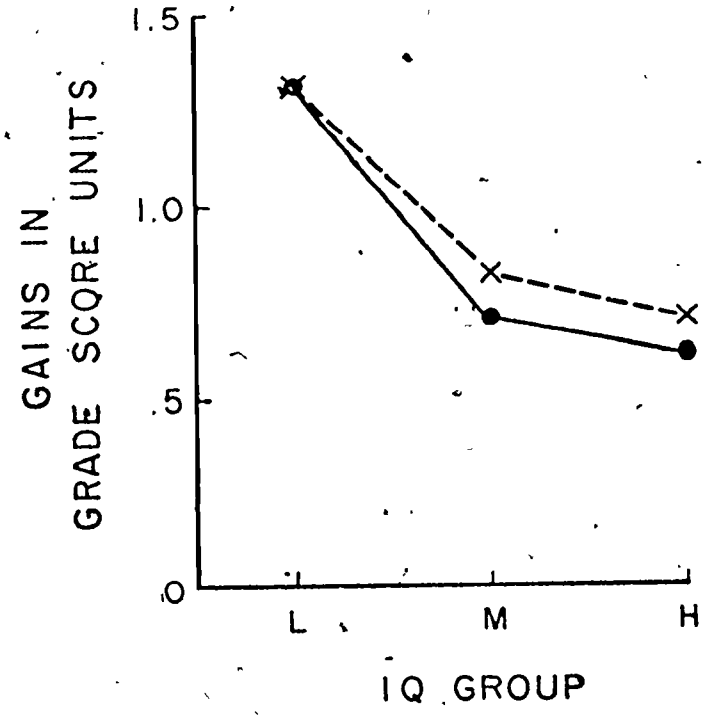
The increasing gap in achievement between different SES or ethnic groups has been interpreted to imply that the schools are differentially effective. The counter part for the illustrative school system is that the system is more effective with high than low ability students. However, there are many reasons why such a conclusion may not be justified. Some of these reasons are discussed below but first the 6th to 7th grade results need to be considered.

Between grades 6 and 7 the mean gains in grade scores on the reading test of the Educational Development Series were: .6 for the high IQ group, .7 for the middle group, and 1.3 for the low group (see Figure 4-1). The pattern is just the reverse of that found for grades 3 to 4. In arithmetic, the grade 6 to 7 pattern was again opposite that of the grade 3 to 4 pattern with gains of .7, .8, and 1.3 for the high, middle, and low ability groups, respectively. Consider the naive interpretation of these data--at grades 3 to 4 the schools might be considered to be more effective with the more able children but at grades 6 to 7 they might be considered to be more effective with the less able. Further, imagine the sort of comparison that might be made among school buildings or among teachers with a predominance of children from different ability levels if the school building mean gains were compared.

In the example just given there are many differences between the data at grades 3 to 4 and those at grades 6 to 7. They are based on tests from different publishers which have different content specifications and different norm groups, and they are based on different types of scales (grade equivalent in one case and grade scores in the other). They also differ in that the same test form spans grades 3 and 4 but different levels which had to be vertically equated were used at grades 6 and 7. These differences may be more than sufficient to explain the seemingly strange results that are shown in Figure 4-1 (Linn, 1974).



(a) 3rd to 4th Grade Gains on the Metropolitan Achievement Tests



(b) 6th to 7th Grade Gains on the Educational Development Series

Reading

Arithmetic

Figure 4-1

Illustration of Average Gains on Two Standardized

Gains in Grade Equivalent Scales. The above results which indicated that students with high pretest scores at grade 3 tended to gain more than their counterparts with low pretest scores may seem contrary to what would be expected from knowledge of correlations of gain scores with pretests. As indicated early in this paper gain scores tend to have a spurious negative correlation with the pretest score. The negative correlation of pretest with gain comes about when the pretest standard deviation is greater than the posttest standard deviation times the correlation between pre and posttests. This will necessarily be the case whenever the pretest and posttest have equal standard deviations. A property of the grade equivalent scale, however, is that the standard deviation of grade equivalent scales tends to increase with grade level and this increase in standard deviation is sufficient to result in a positive correlation between pretest scores and gain scores.

The property of increasing standard deviations for grade equivalent scores at successive grade levels is illustrated by approximating these standard deviations at two grades and for two subtests of three widely used achievement test batteries. The standard deviations were calculated by assuming a normal distribution of grade equivalent scores and subtracting the grade equivalent corresponding to the fiftieth centile from the one corresponding to the eighty-fourth centile. The test batteries that were utilized are the California Achievement Tests (CTB/McGraw Hill, 1970), the Stanford Achievement Tests (Harcourt Brace Jovanovich, 1973) and the Metropolitan Achievement Tests (Harcourt Brace Jovanovich, 1970). For the reading subtests of the three test batteries, the estimated standard deviations for grades two and six for the above tests batteries changed from .925 to 2.27, from 1.70 to 2.65, and from 1.0 to 2.4. The grade two and six standard deviations for the arithmetic subtests of the three batteries changed from .773 to 1.57, from 1.0 to 2.05 and from .7 to 1.4. In general, the estimated standard deviations for grade six are roughly double those for grade two and the necessary condition for a positive correlation between pretest and gain is seen to exist.

Thus, the naive expectation of a gain of one grade equivalent unit in a year's time ignores the positive correlation between gain and pretest that has been observed for the grade equivalent scale. "...normal or typical growth is often defined as one year (1.0) in grade equivalent units for every school year of instruction. However, 1.0 year of growth is typical only for students near the middle of the distribution (Prescott, 1973, p. 55)." As shown by Prescott, by Coleman and Karweit (1970), and by Wrightstone, Hogan and Abbott (undated) students who maintain a constant percentile rank over several years would show average gains that are considerably different than 1.0 when the constant percentile rank deviates substantially from 50.

In order to investigate the generality of the above tendency, the grade equivalent score deviations from grade level for hypothetical students with constant percentile ranks of 20 and of 80 were plotted for several different tests for grades 2 through 6. These results for the reading and arithmetic tests of three widely used achievement test batteries are shown in Figure 4-2. The test batteries for which data are plotted in Figure 4-2 are the Metropolitan Achievement Tests (Harcourt Brace Jovanovich, 1970), the California Achievement Tests (CTB/McGraw, 1970) and the Stanford Achievement Tests (Harcourt Brace Jovanovich, 1973).

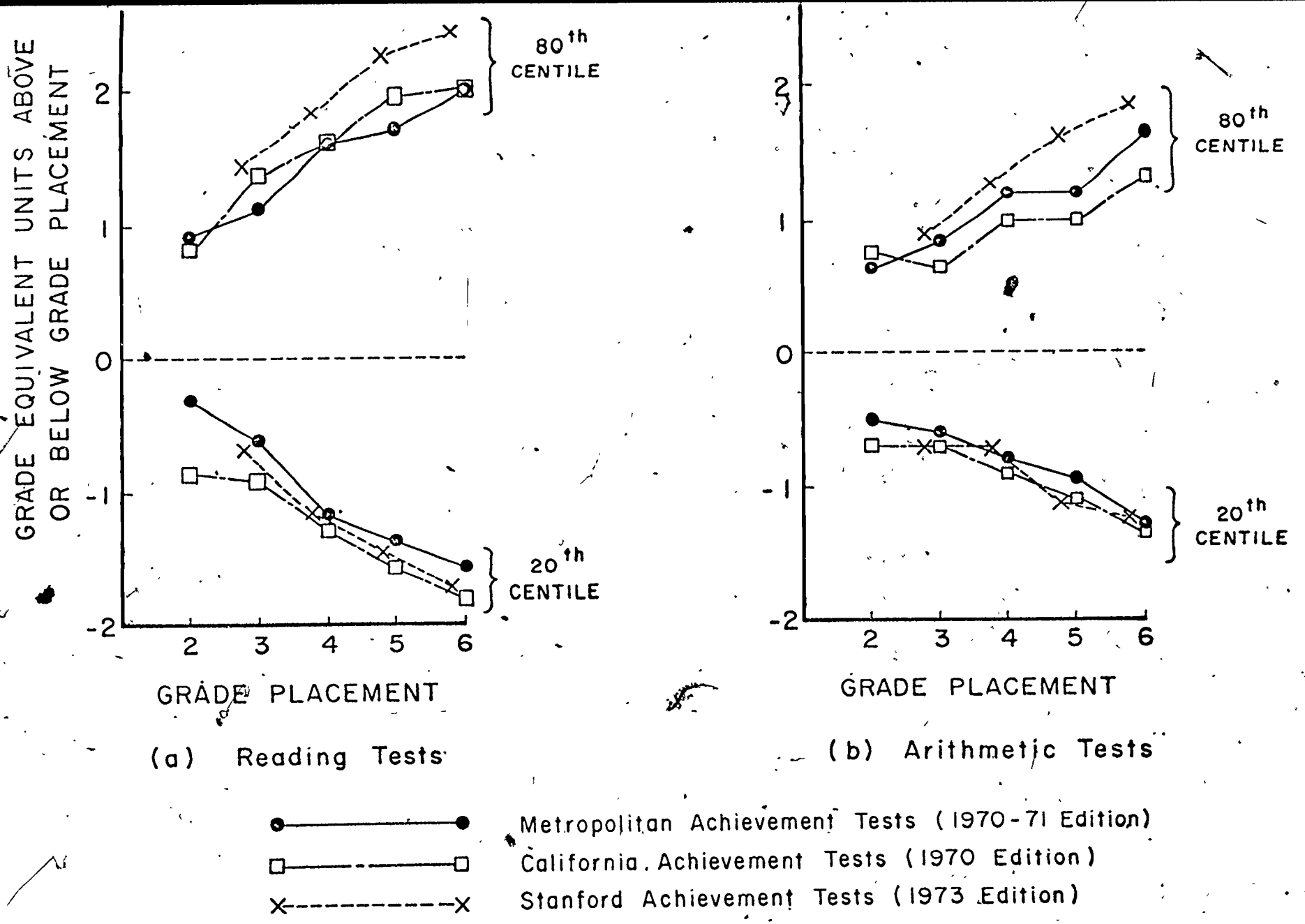


Figure 4-2

Plots of Grade Equivalent Deviations From Grade Placement for the 20th and 80th Centiles on Three Standardized Tests of Reading and of Arithmetic

The graphs shown in Figure 2 provide the basis for several generalizations: (1) the average growth required to maintain a constant percentile rank of 80 is considerably more than 1.0 grade equivalent unit per year, (2) the average growth required to maintain a constant percentile rank of 20 is substantially less than 1.0 grade equivalent unit per year, (3) the average gain in grade equivalent units required to maintain a constant percentile rank of 80 is less for arithmetic tests than for reading tests, and (4) the average gain in grade equivalent units required to maintain a constant percentile rank varies substantially from one test publisher to another.

Based on the results shown in Figure 4-2 the 3rd to 4th grade gains for the illustrative school in Figure 4-1 are quite consistent with what would be expected. The results for grades 3 to 4 certainly are dependent on particular characteristics of the grade equivalent scale that are not really fundamental to notions of student performance. Thus, the result that the more able students tend to gain the most may simply be an artifact of the grade equivalent scale and the naive interpretation that the schools are relatively more effective for high ability than for low ability students is suspect.

A possible conclusion based on the difficulties with the grade equivalent outlined above is that percentile ranks might provide a better scale for comparing growth of groups of students that start at different levels initially. Percentile ranks, however, suffer from other limitations. They tend to spread raw scores out in the middle of the distribution and squeeze them together at the extremes. A distribution of percentile ranks is necessarily rectangular and the raw score distance between the 50th and 55th percentile is much less than the raw score distance between the 90th and 95th percentile. Due to this limitation of percentile ranks, Coleman and Karweith (1970) conclude that they are <u>not</u> a useful type of score for measuring the <u>amount</u> of change but they may be useful for measuring the <u>direction</u> of change.

According to the test manual, the grade scores that were used to summarize the test results for the school system at grades 6 and 7 (Figure 4-1) were "... developed in an attempt to utilize the strong points inherent in percentile rank and grade equivalent norms while minimizing the inherent limitations of such norms scores" (Scholastic Testing Service, 1971, p. 12). Grade scores are obtained from standard scores at each grade level with the mean set equal to the grade placement level and the standard deviation set equal to 1.0. According to the publisher, "Score changes [in grade score units] of more than one unit indicate relatively rapid growth as compared with other pupils; score changes less than one unit indicate relatively slow growth as compared to other students" (Scholastic Testing Service, 1971, p. 13).

A review of grade score scale properties (Linn, 1974) revealed several undesirable characteristics of this type of scale for purposes of measuring change. The most obvious disadvantage of this type of scale is that constant raw scores over several points in time will result in increasing grade scores and "apparent growth." Furthermore, the magnitude of the apparent change varies from one raw score level to another.

As far as the results in section b of Figure 4-1 are concerned, there are two factors which may readily account for the relatively large gains for initially low scoring students and relatively small gains for initially high scoring students. First, by setting the standard deviations at different grade levels equal a negative correlation between pretest and gain is insured. The second factor that is relevant to the particular situation of the grade 6 to 7 results is that different levels of the test were used at grades 6 and 7. As shown by Linn (1974) difficulties in vertically equating tests and the large increase in the scaled score equivalents of minimum and chance level raw scores when the level of the test is changed could easily account for the apparently larger gains of initially low scoring students than their initially high scoring counterparts. Again the results of Figure 4-1 do not provide a basis for generalizations about the relative effectiveness of the school system with different groups of students.

One difficulty with vertically equated tests is the large increase in scaled score equivalents of minimum and chance level raw scores when the level of the test is changed is not limited to grade scores. It is also a potential problem when grade equivalent scores are used with vertically equated tests. Reported in Figure 4-3 for grades 2 through 6 are the grade equivalent scores associated with "chance level" performance on the reading and arithmetic subtests of the three previously used achievement test batteries. As seen from Figure 3, the increase in grade equivalent scores from one level to the next for hypothetical students who respond at random, varies considerably across each publishers' test and across the two subtests. However, even the minimum increase of .6 grade equivalent units would result in apparent growth for students who respond at the chance level.

The Wrong Norms. A number of other difficulties with using norms as standards for evaluating student progress might be mentioned but the illustration of one other problem should suffice. Longitudinal data are often thought to be preferrable to cross-sectional data because of the possibility of cohort differences and because if you are interested in the effects of a school it seems reasonable to look at students who have been in the school for a given period of time. However, the available normative data on standardized achievement tests are cross-sectional. Longitudinal samples often suffer from considerable attrition. Consequently the differences between data for a longitudinal sample and the test norms are apt to be differentially affected by selection factors at different levels. This can be illustrated by data from a national study of academic growth conducted at Educational Testing Service under the direction of Tom Hilton. The data for the following illustration were taken from the extensive set of Tables reported by Hilton and Beaton (1971) and have previously been discussed by Linn (1974).

The longitudinal sample of approximately 3600 students was divided into two groups according to high school curriculum: academic and nonacademic. The scaled score means on one of the tests and the corresponding percentile ranks of the means are plotted in Figure 4-4 for these two groups. The test was the Quantitative Test of the School and College Ability Tests, SCAT (Educational Testing Service, 1957). At the fifth grade the academic group is well above the median of the norm group and the nonacademic group is slightly above the median of the norm group.

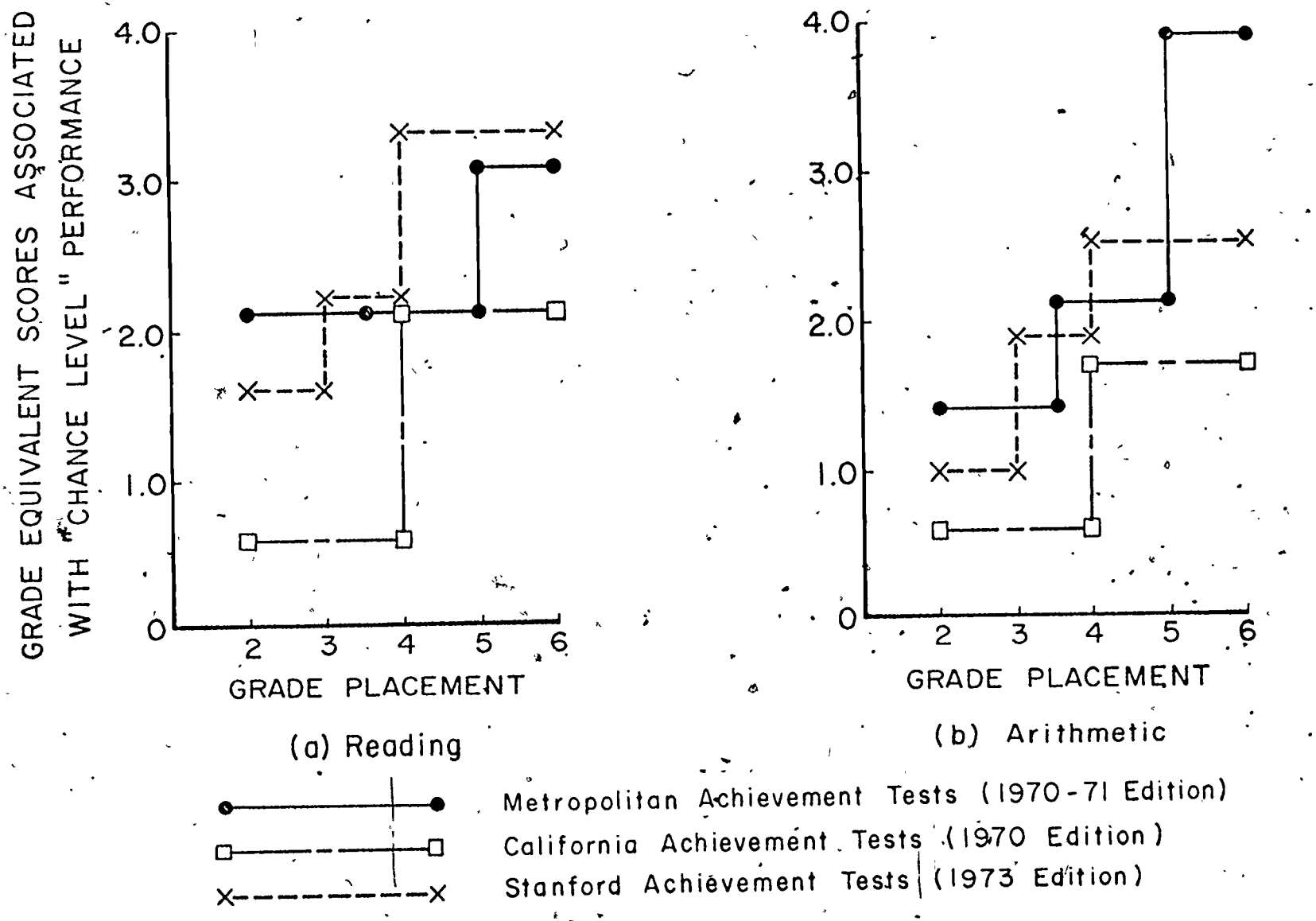


Figure 4-3

Grade Equivalent Scores Associated with "Chance Level" Performance on Three Standardized Tests

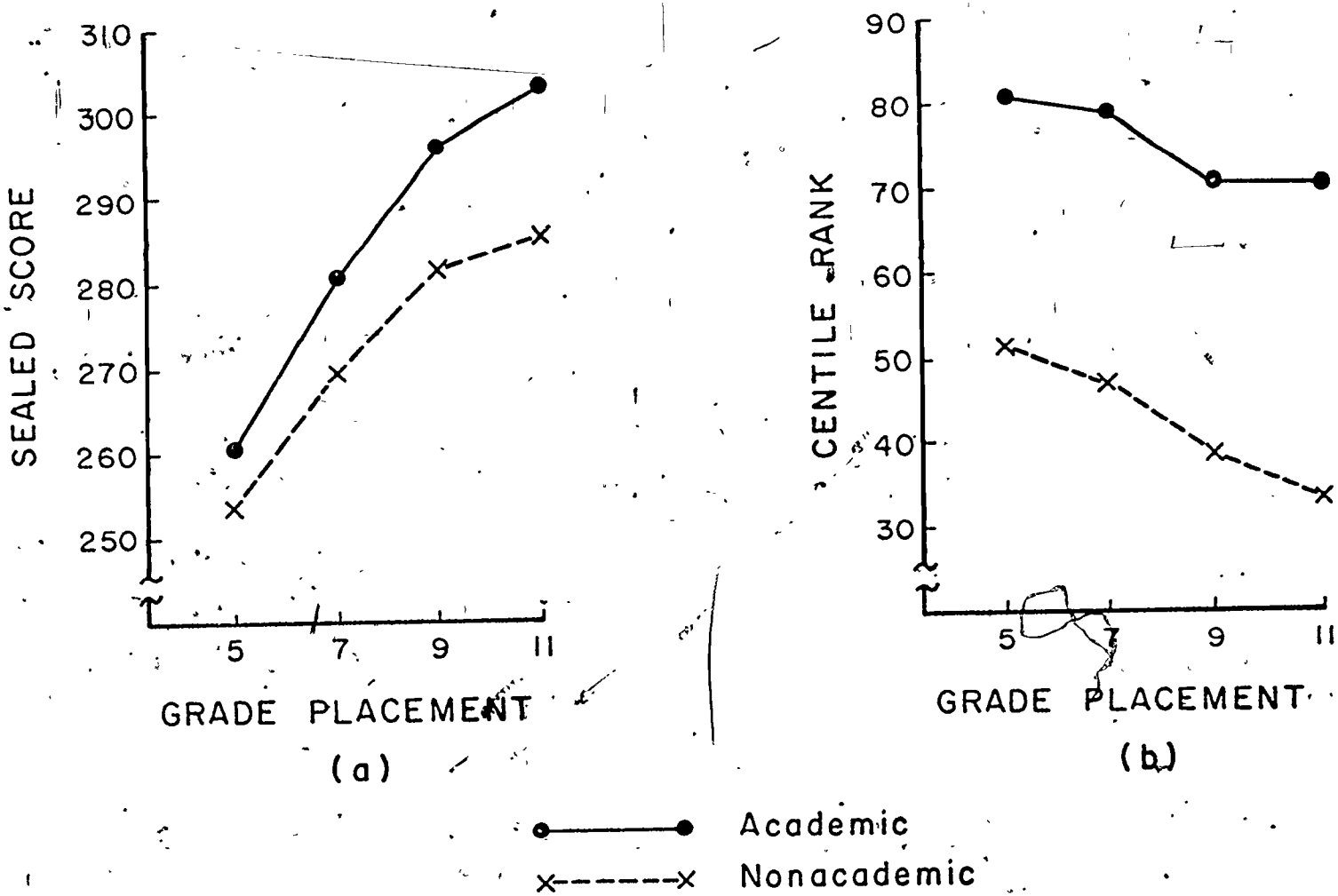


Figure 4-4

Scaled Score Means and Gentile Ranks on SCAT Quantitative Test for ETS Longitudinal

The percentile ranks of the means for both groups drop slightly from grade 5 to grade 7 and more sharply from grade 7 to grade 9. Between grade 9 and grade 11 the academic group maintains about the same percentile rank while the nonacademic group shows another drop.

The initial impressions from Figure 4-4 are that the nonacademic students are falling further and further behind the academic students and both groups of students are losing ground relative to the national norm. Both of these results, however, may be the consequence of a common problem encountered in longitudinal studies, namely attrition. The initial ETS growth study consisted of about 9,000 5th grade students. Only about 40 percent of these students had test score data at grades 5, 7, 9, and 11 and the nonrandom nature of the attrition is apt to have different implications at 5th grade than at 11th grade. For example, students who drop out of school between the 5th and 11th grades are available for the norms group at grade 5, but not at grade 11. For the longitudinal sample they are excluded at both points in time (Linn, 1974).

Problems due to using cross-sectional norms can arise even where the longitudinal data cover only two points in time within a single school year. Data from two points in a single year usually do not have a major attrition problem such as was encountered for the data in Figure 4-4. Nonetheless, using fall data to interpolate the norms for other points in the year may result in misleading "growth expectations." For example, Beck (1975) has recently shown that norms based only on fall testing tend to underestimate the actual spring performance of a longitudinal sample that is tested in the fall and again in the spring.

Regression Approaches to Accountability

One of the better known approaches to developing an accountability system is the one proposed by Dyer (1970; Dyer, Linn & Patton, 1969). His approach, which was first described before the term "accountability" came into popular use (Dyer, 1966), is based on what he calls "the pupil-change model of a school." Actually student change per se is never assessed in Dyer's approach, instead, regression equations are used to compute residual mean performance for a school. These residuals form the basis for obtaining "school effectiveness indices."

As initially conceived, the Dyer approach would distinguish four major categories of variables called input, surrounding conditions, educational process, and output. The input and output categories of variables refer to student characteristics measured before and after a given period of schooling. While these groups of variables were broadly conceived to include a wide array of measures, as implemented the input category is apt to consist of pretest scores and the output of posttest scores.

Surrounding condition variables consist of the variety of home, school and community characteristics that describe the conditions within which the school operates. Dyer (1970) distinguishes between surrounding condition variables that are relatively "hard to change" and those that are relatively "easy to change." Finally, the education process variables consist of activities of the school that may influence student achievement.

With the four categories of variables in hand regression analyses involving the input, output and hard to change surrounding conditions would be used to obtain "school effectiveness indices" for each output measure. Specifically, using school means, a given output or posttest score would be regressed on the input measures and hard to change surrounding conditions would be used to obtain "school effectiveness indices" for each output measure. Specifically, using school means, a given output or posttest score would be regressed on the input measures and hard to change surrounding condition variables. Schools with observed mean scores on the posttest that were above the value predicted for that school would receive relatively high school effectiveness indices. Schools with posttest means lower than predicted would receive relatively low indices.

Only after the school effectiveness indices are obtained would the easy to change surrounding conditions and the school process variables come into play. The focus would be on outliers, i.e., those schools that have posttest means much better (or worse) than predicted from the pretests and hard to change surrounding conditions. The extreme outliers, which in another context would be called "overachievers and underachievers" (Thorndike, 1963), would then be compared in terms of the easy to change surrounding condition variables and the educational process variables.

Dyer was well aware that his proposed approach gives no guarantee of finding the characteristics of schools that produce the maximum achievement. Rather the approach was conceived of as a kind of search strategy for identifying variables that might be instrumental to better student performance. The actual efficacy of these variables could then be investigated in experimental studies.

There are a number of questions that might be raised concerning Dyer's approach. As indicated in the first section of this paper, residuals still may be questionable. Dyer, Linn and Patt[illegible] (1969) provided results that are relevant to one type of reliability of the school residuals. School systems were subdivided into two random halves and residuals computed for each half sample. The correlations of the half sample residual scores ranged from .73 to .88 for six different posttests. While these results suggest reasonable stability, less encouraging results were obtained by Forsyth (1973) when he investigated another type of reliability.

Forsyth (1973) obtained school residuals according to the Dyer model for two successive time intervals (posttests obtained in 1968 and in 1969). The correlations between residuals obtained for schools at the two different points in time ranged from .11 to .50 for 10 posttests with a median correlation of only .28. Thus, it would appear that the residuals may be relatively stable for one subsample of students to another within a single year but relatively unstable from one year to the next. This instability over time is seen as a major limitation on the potential usefulness of this approach.

Recently, Marco (1974) compared four different methods of obtaining school effectiveness indices in addition to the one originally suggested

by Dyer. He found that all five methods yielded indices that were highly intercorrelated and relatively stable from one half sample to another. His study does not address the issue of stability over time or the practical utility of the indices, however.

CONCLUSION

This paper has ranged over a fairly broad spectrum of topics that share as a common thread concern about measuring change from pre to posttesting periods. Problems in measuring change abound and the virtues in doing so are hard to find. Major disadvantages in the use of change scores are that they tend to conceal conceptual difficulties and they give misleading results. The former tendency is apparent when change scores are used to compare preexisting groups which tends to conceal to the arbitrariness of this particular form of adjustment. The latter tendency is apparent where various standardized test scales such as grade equivalents or percentile ranks are used to assess gains of different groups of students.

To conclude with Cronbach and Furby (1970) "...that investigators who ask questions regarding gain scores would ordinarily be better advised to frame their questions in other ways (p. 80)" may seem very discouraging. If so, however, it is probably because more is expected from gain scores than they can reasonably be expected to provide. They cannot, for instance, be expected to make up for the lack of random assignment, nor can other adjustment techniques. For most purposes, a pretest score is best treated on the same footing as other measures that are obtained at the time of the pretest. Where appropriate, regression analyses that treat the pretest no differently than other independent variables (or predictors) and the posttest as the dependent variable avoids many of the difficulties that are introduced by gain scores.

REFERENCES

American Psychological Association, Standards for Educational and Psychological Tests, Washington, D. C.: 1974.

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement, 2nd Edition, Washington, D. C.: American Council on Education, 1971.

Beck, M. D., Development of empirical "growth expectancies" for the Metropolitan Achievement Tests, Presented at the meeting of the National Council on Measurement in Education, Washington, D. C., 1975.

Bereiter, C. Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.) Problems in Measuring Change. Madison: University of Wisconsin Press, 1963, pp. 3-20.

Coleman, J. S. & Karweit, N. L. Measures of School Performance. Santa Monica, California: Rand, R-488-RC, July 1970.

Cronbach, L. J. & Furby, L. How we should measure "change" --or should we? Psychological Bulletin, 1970, 74, 68-80.

CTB/McGraw-Hill, California Achievement Tests (1970 ed.). Monterey, California: CTB/McGraw-Hill, 1970.

DuBois, P. H. Multivariate Correlational Analysis. New York: Harper, 1957.

Dyer, H. S. The Pennsylvania Plan: Science Foundation, 1966, 50, 242-248.

Dyer, H. S. Toward objective criteria of professional accountability in the schools of New York City. Phi Delta Kappan, 1970, 52, 206-211.

Dyer, H. S., Linn, R. L. & Patton, M. J. A comparison of four methods of obtaining discrepancy measures based on observed and predicted school system means on achievement tests. American Educational Research Journal, 1969, 6, 591-605.

Educational Testing Service. School and College Ability Test. Princeton, New Jersey, Educational Testing Service, 1957.

Elashoff, J. D. Analysis of covariance: a delicate instrument. American Educational Research Journal, 1969, 6, 383-402.

Feldt, L. S. A comparison of the precision of three experimental designs employing a concomitant variable. Psychometrika, 1958, 23, 335-353.

Forsyth, R. A. Some empirical results related to the stability of performance indicators in Dyer's student change model of an educational system. Journal of Educational Measurement, 1973, 10, 7-12.

Glass, G. V. The many faces of "educational accountability". Phi Delta Kappan, 1972, 53, 636-639.

Goulet, L. R. Longitudinal and time-lag designs in educational research: an alternate sampling model, Review of Educational Research, in press.

Harcourt Brace Jovanovich, Metropolitan Achievement Tests (1970 ed.), New York: Harcourt Brace Jovanovich, 1970.

Harcourt Brace Jovanovich, Stanford Achievement Tests (1973 ed.). New York: Harcourt Brace Jovanovich, 1973.

Hilton, T. L. & Beaton, A. E. Stability and instability in academic growth -- a compilation of longitudinal data. Final Report, August, 1971, Educational Testing Service, Grant No. OEG-2-7000013(509), U. S. Office of Education.

Hilton, T. L. & Patrick, C. Cross-sectional versus longitudinal data: An empirical comparison of mean differences in academic growth. Journal of Educational Measurement, 1970, 7, 15-24.

Lindquist, E. F. & Hieronymus, A. N. Manual for administrators, supervisors and counselors Iowa Tests of Basic Skills. Boston, Massachusetts: Houghton Mifflin Company, 1964.

Linn, R. L. The use of standardized test scales to measure growth. Conference on Policy Research: Methods and Implications. University of Wisconsin, Madison, Wisconsin, May 1974.

Linn, R. L. & Werts, C. E. Errors of inference due to errors of measurement, Educational and Psychological Measurement, 1973, 33, 531-543.

Lord, F. M. The measurement of growth. Educational and Psychological Measurement, 1956, 16, 421-437. See also Errata, ibid., 1957, 17, 452.

Lord, F. M. Further problems in the measurement of change. Educational and Psychological Measurement, 1958, 18, 437-454.

Lord, F. M. Large sample covariance analysis when the control variable is fallible. Journal of the American Statistical Association, 1960, 55, 309-321.

Lord, F. M. Elementary models for measuring change. In C. W. Harris (Ed.) Problems in Measuring Change. Madison: University of Wisconsin Press, 1963, 21-38.

Lord, F. M. A paradox in the interpretation of group comparisons, Psychological Bulletin, 1967, 68, 304-305.

Lord, F. M. Statistical adjustments when comparing pre-existing groups. Psychological Bulletin, 1969, 72, 336-337.

Manning, W. H. & DuBois, P. H. Correlational methods in research on human subjects. Perceptual Motor Skills, 1962, 15, 287-321.

Marco, G. L. A comparison of selected school effectiveness measures based on longitudinal data. Journal of Educational Measurement, 1974, 11, 225-234.

Marks, E. & Martin, C. G. Further comments relating to the measurement of change. American Educational Research Journal, 1973, 10, 179-191.

McNemar, Q. On growth measurement. Educational and Psychological Measurement, 1958, 18, 47-55.

Meehl, P. E. Nuisance variables and the ex post facto design. In M. Radner & S. Winokur, (Eds.) Minnesota Studies in Philosophy of Science Volume 4, Minneapolis: University of Minnesota Press, 1970.

O'Connor, E. F., Jr. Extending classical test theory to the measurement of change. Review of Educational Research, 1972, 42, 73-97.

Porter, A. C. The effects of using fallible variables in the analysis of covariance. (Doctoral dissertation, University of Wisconsin), Ann Arbor, Michigan: University Microfilms, 1967, No. 67-12, 147.

Prescott, G. A. Manual for Interpreting: Metropolitan Achievement Test, New York: Harcourt Brace Jovanovich, Inc., 1973.

Rosenshine, B. & McGaw, B. Issues in assessing teacher accountability in public education. Phi Delta Kappan, 1972, 53, 640-643.

Rubin, D. Estimating Causal Effects of Treatments in Experimental and Observational Studies. Princeton, N. J.: Educational Testing Service, Research Bulletin, 72-39, 1972.

Scholastic Testing Service. Educational Development Series Technical Report - Elementary Level - Spring 1971. Bensenville, Illinois: Scholastic Testing Service, 1971.

Stake, R. E. Testing hazards in performance contracting. Phi Delta Kappan, 1971, 52, 583-589.

Tatsuoka, K. K. Vector-Geometric and Hilbert-Space Reformulations of Classical Test Theory, Doctoral Dissertation, University of Illinois, 1975.

Thorndike, R. L. The concepts of over- and underachievement. New York: Columbia University, Teachers College, Bureau of Publications, 1963.

Thorndike, R. L. Intellectual status and intellectual growth. Journal of Educational Psychology, 1966, 57, 121-127.

Wargo, M. J. et al. ESEA Title I: A reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970. Palo Alto, California: American Institutes for Research, 1972.

Werts, C. E., Joreskog, K. G. & Linn, R. L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972, 32, 655-678.

Werts, C. E. & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73, 17-22.

Werts, C. E. & Linn, R. L. Analyzing school effects: ANCOVA with a fallible covariate, Educational and Psychological Measurement, 1971, 31, 95-104.

Wrightstone, J. W., Hogan, T. P., & Abbott, M. M. Accountability in education and associated measurement problems. Test Service Notebook 33, New York: Harcourt Brace Jovanovich, Inc., (undated).

CHAPTER 5

VERTICALLY EQUATED TEST FORMS

In large scale testing programs it is frequently necessary and desirable to have several forms of a test. Multiple forms are essential for admissions tests such as the College Board's Scholastic Aptitude Test or the American College Testing Program's Tests. The purpose of the equating is to convert the raw scores obtained from two forms of the test "...so that scores derived from the two forms after conversion will be directly equivalent (Angoff, 1971, p. 562)" In the case of admissions tests, equating is essential because comparisons are made between persons who take different forms of the test and without the equating persons who happened to take one form of the test that was inadvertently more difficult than another form would be at a disadvantage relative to their peers who happened to take the easier form.

Equating test forms that are designed to measure the same thing for the same population is sometimes referred to as horizontal equating (see, for example, Educational Testing Service, 1957, pp. 7-9). Vertical equating, on the other hand refers to the process of converting scores of forms of a test designed for populations at different educational levels to a single scale. In horizontal equating, different forms of the test would normally be designed to have comparable item content and similar distributions of item statistics. The equating adjusts for unintended differences in difficulty of the tests or differences in distributions of the examinees. In contrast, forms to be vertically equated differ intentionally in the difficulty of the items for a single population of examinees and in their content specifications as well. For example, an appropriate arithmetic item might be $4 + 3 = ?$ at grade 1, $155 - 62 = ?$ at grade 3, $67 \times 4 = ?$ at grade 5, and $5.45 \div .25 = ?$ at grade 7. To be sure, such items are all in the general domain of arithmetic but they are not necessarily indicators of a single common trait. In other achievement areas even greater diversity of item type, difficulty, and content frequently can be found as changes in the level of a test occur while a common name and supposedly common scale is maintained. It is no surprise that the problem of vertical equating is substantially more difficult than that of horizontal equating.

In this section, the two most commonly used equating procedures will be briefly reviewed. The adequacy of these methods for the vertical equating problem will then be considered. Finally, consideration will be given to alternative equating methods with special emphasis on the use of the Rasch model.

LINEAR AND EQUIPERCENTILE METHODS

"Two scores, one on form X and the other on form Y (where X and Y measure the same function with the same degree of reliability), may be considered equivalent if their corresponding percentile ranks in any given group are equal (Angoff, 1971, p. 563)." This commonly accepted definition suggests immediately the equipercentile method of equating. All that is required for the equipercentile method of equating is the

cumulative frequency distribution for each test. The k^{th} score level on form X, X_k, is converted to the same scaled score as the 1^{th} score level of test Y, Y_1, if the percentiles of X_k and Y_1 are the same. In practice, smoothed frequency distributions are typically used and raw scores on the tests corresponding to some predetermined set of percentile ranks are found by interpolation. Also, there are a variety of different study designs that might be used for the equating. For example, both tests may be administered to a single group, the tests may be administered to a different random sample from the same population, or the tests along with a common anchor test may be administered to a sample from different populations. For a detailed description of these and other possible designs see Angoff, (1971). Ignoring these procedural details, however, the equipercentile method is quite straight forward.

Linear equating would assign the same scaled score to scores X_k and Y_1 if they correspond to the same standard score, that is if

$$\frac{X_k - \bar{X}}{S_x} = \frac{Y_1 - \bar{Y}}{S_y}$$

where $\bar{X}$, $\bar{Y}$, S_x, and S_y are the means and standard deviations of X and Y respectively. As noted by Angoff (1971), the equipercentile and linear equating methods coincide if the two marginal distributions differ only in their first and second moments. More generally, the two methods will yield similar results when the raw score frequency distributions are similar.

For purposes of vertical equating there are two important aspects of the above paragraphs that need to be considered. (1) Linear equating might be expected to be less adequate than equipercentile equating for the vertical situation because there is less reason to expect X and Y to have distributions of about the same shape. (2) A key aspect in the definition of equivalent scores given above is the requirement that the percentile ranks be equal "...in any given group...". If this requirement is not met then the conversion will not be unique. More will be said about this second point below but first a few comments are offered regarding the likely utility of the linear method in vertical equating.

THE ANCHOR TEST STUDY

Undoubtedly the largest equating study ever conducted was the Anchor Test Study (Bianchini and Loret, 1974). (For a more complete review of the Anchor Test Study see Appendix A.) This study, and its supplement equated eight widely used standardized reading tests at

grades 4, 5, and 6. Although the equating was done separately within each grade, and thus the equating might naturally be viewed as horizontal, the results are in fact quite relevant to the problem of vertical equating. The tests being equated differed substantially in difficulty level as well as in content specifications. Furthermore, there were a variety of patterns of common versus different forms used at grades 4, 5, and 6 which make it possible to compare equated scores at one grade level with those at another.

The various pairs of tests involved in the anchor test study were equated by both the equipercentile and linear methods. These methods were compared in terms of the estimated errors of equating which were obtained by the use of McCarthy's balanced half-sample replication method (1966). The equating design consisted of a set of eight balanced half-samples. These half-sample replications were used to compute the root-mean squared deviation of equivalent scores on the anchor test for each half-sample replication about the anchor test equivalent scores for the full sample. Based on the estimated errors the equipercentile method was judged to be clearly superior to the linear method. Furthermore, the degree of superiority was greatest for those tests which differed most from the anchor test in their level of difficulty. Based on these results and logical considerations about the likelihood that distributions of forms to be vertically equated will differ in moments higher than the second, the equipercentile method seems preferable to the linear method in the vertical situation.

The Anchor Test Study also provides another form of evidence that is relevant for the problem of vertical equating. Two tests involved in the study changed levels between grades 4 and 5, three tests changed levels between grades 5 and 6, two tests involved a single level over all three grades and one test changed levels at each grade. These different patterns of levels make possible a variety of comparisons of the equatings of two levels of one test to a single level of another test. For example, the same level of California Achievement Tests, CAT, (CTB, McGraw-Hill, 1970) was used at grades 4 and 5 but different levels of the Metropolitan Achievement Tests, MAT, (Harcourt, Brace, Jovanovich, 1970) were used at those grades. By using the CAT equivalencies of the MAT, it is possible to convert the MAT Elementary Level Reading scores to equivalent Intermediate Level Reading scores. For purposes of illustration, a few scores of the CAT at grade 4 were selected and the equivalent Elementary Level MAT scores were noted. The same CAT scores were then used at grade 5 to find the equivalent Intermediate Level MAT raw scores. These scores are shown in Table 5-1. The publisher's norms were used to convert the equated MAT Elementary and Intermediate raw scores to grade equivalent scores. The resulting grade equivalent scores are also reported in Table 5-1. Finally, the grade equivalent score at grade 4 was subtracted from the corresponding score at grade 5 and the difference was recorded in the last column of Table 5-1.

If the two columns of grade equivalent scores in Table 5-1 are compared, some non-trivial differences in the grade equivalents can be observed. The largest of the differences in corresponding grade equivalents shown in Table 5-1 occurs for MAT raw scores that are equivalent to a CAT raw score of 60. At this level, the grade equivalent scores are 6.6 at grade 4 and 7.4 at grade 5 for a difference of 0.8 grade equivalent units which would presumably be interpreted as almost a "year's gain." Except at the extremely high end of the distribution, the grade equivalents tend to be larger at grade 5 than at grade 4.

A number of other test combinations could be used to produce tables such as Table 5-1. For example, the grade 4 and grade 5 MAT scores could be equated through their links to the Comprehensive Tests of Basic Skills, CTBS, (CTB, McGraw-Hill, 1968) rather than through the CAT. This was done and the results are reported in Table 5-2. As can be seen in Table 5-2, the grade equivalents at grade 5 again tend to be higher than the corresponding grade equivalents at grade 4.

The results in Tables 5-1 and 5-2 suggest that changes in grade equivalent units might differ substantially depending on whether a single level of a test or two vertically equated levels of a test are being used in, say, a longitudinal research study. In particular, larger gains would be expected using the Elementary level of the MAT at grade 4 and the Intermediate level of the MAT at grade 5 than would be expected if either level 2 of the CTBS or level 3 of the CAT were used at the two grades.

In addition to the grade equivalent scores, vertically equated "standard scores" were also compared. The standard scores reported by the test publisher of the MAT test are scaled to range from grade 1 to grade 9. At grade 4, the mean scaled score is about 66 and the associated standard deviation is about 14. By grade 9, the mean and standard deviation are approximately 96 and 17 respectively.

The grade 4 and grade 5 standard scores of the MAT were compared by converting equivalent raw scores on the Elementary and Intermediate Levels of the MAT to standard scores. When the CAT was used to define equivalent raw scores on the MAT, the results in Table 5-3 were obtained. The results in Table 5-4 were obtained by using the CTBS to define equivalent MAT raw scores for the two levels of the MAT. For all but relatively high scores, the Intermediate Level MAT standard scores are somewhat higher than the "equated" Elementary Level standard scores. This is true whether the equating is accomplished via the CAT (Table 5-3) or via the CTBS (Table 5-4). Furthermore, the magnitude of the difference in standard scores is relatively large in some parts of the score distribution.

It might be noted that the largest differences in standard scores reported in Tables 5-3 and 5-4 occur at the extremes where relatively few observations are expected. Even in the central part of the score range, however, the differences are as large as a third of a within grade standard deviation. A difference as big as a third of a standard deviation is apt to loom large relative to the magnitude of "effects" that are being evaluated. Thus, whether grade equivalent scores or other

TABLE 5-1

Total Reading Equivalent Scores on the MAT Elementary and Intermediate Levels (Grade Equivalents via CAT)

Level 3 CAT Raw Scores (Grades 4 & 5)	Equivalent MAT Raw Scores and Corresponding Grade Equivalents: Elementary Level (Grade 4) Raw	Elementary Level (Grade 4) GE	Intermediate Level (Grade 5) Raw	Intermediate Level (Grade 5) GE	Difference in GE Scores (Grade 5 minus Grade 4)
80	94	9.9	91	9.8	-0.1
70	89	8.4	76	8.4	0.0
60	84	6.6	63	7.4	0.8
50	76	5.2	51	5.5	0.3
40	63	3.7	39	4.4	0.7
30	45	3.2	29	3.5	0.3
20	26	2.3	20	2.6	0.3
10	12	1.3	8	1.4	0.1

TABLE 5-2

Total Reading Equivalent Scores on the MAT Elementary and Intermediate Levels (Grade Equivalents via CTBS)

Level 2 CTBS Raw Scores (Grades 4 & 5)	Equivalent MAT Raw Scores and Corresponding Grade Equivalents				Difference in GE Scores (Grade 5 minus Grade 4)
	Elementary Level (Grade 4)		Intermediate Level (Grade 5)		
	Raw	GE	Raw	GE	
80	93	9.8	87	9.8	0.0
70	86	7.3	69	6.9	-0.4
60	78	5.4	55	5.7	0.3
50	68	4.3	44	4.9	0.6
40	56	3.5	35	4.2	0.7
30	41	2.9	28	3.5	0.6
20	24	2.0	20	2.6	0.6
10	12	1.3	10	1.6	0.3

TABLE 5-3

Total Reading Equivalent Scores on the MAT Elementary and Intermediate Levels (Scaled Scores* via CAT)

Level 3 CAT Raw Scores (Grades 4 & 5)	Equivalent MAT Raw Scores and Corresponding Scaled Scores: Elementary Level (Grade 4)		Intermediate Level (Grade 5)		Difference in Scaled Scores (Grade 5 minus Grade 4)
	Raw	Scaled	Raw	Scaled	
80	94	119	91	117	-2
70	89	94	76	91	-3
60	84	84	63	83	-1
50	76	75	51	77	2
40	63	66	39	70	4
30	45	58	29	62	4
20	26	47	20	52	5
10	12	26	8	29	3

*MAT Standard Scores

TABLE 5-4

Total Reading Equivalent Scores on the MAT Elementary and Intermediate Levels (Scaled Scores* via CTBS)

Level 2 CTBS Raw Scores (Grades 4 & 5)	Equivalent MAT Raw Scores and Corresponding Scaled Scores: Elementary Level (Grade 4) Raw	Elementary Level (Grade 4) Scaled	Intermediate Level (Grade 5) Raw	Intermediate Level (Grade 5) Scaled	Difference in Scaled Scores (Grade 5 minus Grade 4)
80	93	112	87	105	-7
70	86	88	69	86	-2
60	78	77	55	79	2
50	68	69	44	73	4
40	56	62	35	67	5
30	41	56	28	61	5
20	24	45	20	52	7
10	12	26	10	34	8

*MAT Standard Scores

scaled scores are used, change observed on the same level of a test is apt to yield different results than change observed over two vertically equated levels of a test.

In Tables 5-1 through 5-4, except for very high scores, there is a consistent tendency for the "equated" scores based on the higher level form to be larger than their counterparts based on the lower level form. If this was a general trend, then it might be possible to compensate for the tendency. Unfortunately, this trend does not hold for all test combinations.

Additional comparisons of vertically equated scores based on the results of the Anchor Test Study are reported in Tables 5-5 through 5-8. The results in Tables 5-5 through 5-8 provide comparisons of results for grades 5 and 6. At those grades, the same level (Intermediate II) of the Stanford Achievement Tests, SAT, (Harcourt, Brace, Jovanovich, 1973) was used while different levels of the CAT (Levels 3 and 4) and of the CTBS (Levels 2 and 3) were used. In Table 5-5, selected raw scores on the SAT are reported along with equivalent CAT Level 3 and CAT Level 4 raw scores and associated grade equivalent scores. The differences in "equated" grade equivalent scores are also reported in Table 5-5. A similar set of results for CTBS Level 2 and Level 3 grade equivalent scores are reported in Table 5-7. The results in Tables 5-6 and 5-8 were obtained in parallel fashion except that other vertically equated scaled scores that are reported by the publisher are used.

The results for the CAT grade equivalent scores (Table 5-5) have a pattern just the opposite of the one previously encountered for the MAT. That is, except for the highest scores, the higher level form tends to yield lower grade equivalent scores than the "equated" score of the lower level form. It should also be noted that the magnitude of the grade equivalent score differences in Table 5-5 tend to be smaller for scores in the middle of the range than were the differences in Tables 5-1 or 5-2.

The results in Table 5-6 are based on the CAT Achievement Development Scale Scores. These scores are scaled to span grades 1 to 12 with a range of scores from 100 to 900. The mean at grade 10 is set at 600 and the standard deviation at 100. At grade 4, the mean is about 400 and the standard deviation about 65. The results in Table 5-6 are similar to those in Table 5-5. The Achievement Development Scale Scores are lower for Level 4 than for Level 3 except at the very high end of the score distribution. The magnitude of the difference for the middle range of scores is only about an eighth of a within grade standard deviation or less.

In Table 5-7, the CTBS Level 2 and Level 3 grade equivalent scores that correspond to common SAT scores are reported. In the middle part of the score range, the Level 2 grade equivalents are higher than their Level 3 counterparts and the opposite is true at both extremes of the score distribution. The magnitude of the difference in the middle part of the score distribution is 0.3 or 0.4 grade equivalent units. Similar

TABLE 5-5

Total Reading Equivalent Scores on the CAT
Level 3 and Level 4 (Grade Equivalents via SAT)

Intermediate II SAT Raw Scores (Grades 5 & 6)	Equivalent CAT Raw Scores and Corresponding Grade Equivalents				Difference in GE Scores (Grade 6 minus Grade 5)
	Level 3 (Grade 5)		Level 4 (Grade 6)		
	Raw	GE	Raw	GE	
110	82	12.9	82	13.6	0.7
100	80	11.4	71	11.5	0.1
90	77	10.1	62	9.8	-0.3
80	72	8.5	55	8.5	0.0
70	68	7.7	48	7.5	-0.2
60	63	7.0	42	6.8	-0.2
50	56	6.1	36	5.9	-0.2
40	47	5.1	29	4.9	-0.2
30	35	3.9	22	3.5	-0.4
20	22	2.4	16	2.2	-0.2
10	13	1.1	9	0.6	-0.5

TABLE 5-6

Total Reading Equivalent Scores on the CAT
Level 3 and Level 4 (Scaled Scores* via SAT)

Intermediate II SAT Raw Scores (Grades 5 & 6)	Equivalent CAT Raw Scores and Corresponding Scaled Scores				Difference in Scaled Scores (Grade 6 minus Grade 5)
	Level 3 (Grade 5)		Level 4 (Grade 6)		
	Raw	Scaled	Raw	Scaled	
110	82	665	82	757	92
100	80	625	71	626	1
90	77	580	62	566	-14
80	72	530	55	528	- 2
70	68	503	48	497	- 6
60	63	480	42	474	- 6
50	56	454	36	450	- 4
40	47	424	29	415	- 9
30	35	380	22	364	-16
20	22	318	16	306	-12
10	13	259	9	232	-27

*CAT Achievement Development Scale Scores

TABLE 5-7

Total Reading Equivalent Scores on the CTBS Level 2 and Level 3 (Grade Equivalent via SAT)

Intermediate II SAT Raw Scores (Grades 5 & 6)	Equivalent CTBS Raw Scores and Corresponding Grade Equivalents: Level 2 (Grade 5)		Level 3 (Grade 6)		Difference in GE Scores (Grade 6 minus Grade 5)
	Raw	GE	Raw	GE	
110	85	11.9	84	12.9	1.0
100	82	11.5	75	11.5	0.0
90	79	9.7	66	9.4	-0.3
80	76	8.7	59	8.3	-0.4
70	72	7.6	52	7.3	-0.3
60	68	6.9	45	6.5	-0.4
50	62	6.0	37	5.6	-0.4
40	53	5.1	30	4.7	-0.4
30	38	3.9	22	3.6	-0.3
20	23	2.7	16	2.5	-0.2
10	12	1.2	9	2.0	0.8

TABLE 5-8

Total Reading Equivalent Scores on the CTBS Level 2 and Level 3 (Scaled Scores* via SAT)

Intermediate II SAT Raw Scores (Grades 5 & 6)	Equivalent CTBS Raw Scores and Corresponding Scaled Scores: Level 2 (Grade 5)		Level 3 (Grade 6)		Difference in Scaled Scores (Grade 6 minus Grade 5)
	Raw	Scaled	Raw	Scaled	
110	85	744	84	786	42
100	82	660	75	641	-19
90	79	612	66	579	-33
80	76	554	59	543	-11
70	72	523	52	513	-10
60	68	497	45	483	-14
50	62	465	37	451	-14
40	53	433	30	421	-12
30	38	386	22	370	-16
20	23	325	16	314	-11
10	12	236	9	247	-11

*CTBS Expanded Standard Scores

results are reported in Table 5-8 using the CTBS Expanded Standard Scores which range from 100 to 900 with a mean and standard deviation at grade 10 of 600 and 100 respectively. The magnitude of the differences in Table 5-8 tends to be about one fifth of the standard deviation observed at grade 5 (which is about 72).

In summary, the results in Tables 5-1 through 5-8 raise doubts about the adequacy of the vertical equating. Change observed on a single level of a test is apt to have a different meaning than the same change observed on vertically equated levels of the same test. Unfortunately, the direction of the difference is apparently not consistent.

THE RASCH MODEL

An important aspect of the definition of equivalent scores that was mentioned above is that the corresponding percentile ranks be equal for "any given group." With presently used methods of equating, this ideal is only roughly approximated for vertically equated test forms. This may simply be a reflection of the difficulty of the task rather than a fault of the methods. It is possible, however, that a rather different approach to the problem would yield better results. If so, that would be a valuable contribution to longitudinal research studies. An approach that appears particularly promising for the problem of vertical equating is one based on the Rasch (1960, 1966a, 1966b) model.

The appeal of the Rasch model is apparent in Wright's (1968) description of the model as providing "person-free test calibration" and "item-free person measurement." What is meant by person-free test calibration is that the item parameters that are estimated are invariant for all groups of persons. Item-free person measurement, on the other hand, means that once items have been calibrated that except for errors of measurement, the same score would be obtained for an individual regardless of which subset of items is used for the measurement. These properties are precisely what is needed for the vertical equating problem.

Rasch's model is a particular instance of a latent trait model and presumably the comments about the potential use of the model in achieving invariant item parameter and person scores could apply to other latent trait models. The primary potential advantage of the Rasch model is its relative simplicity in that items are characterized by a single parameter. This characteristic may at the same time be the primary potential disadvantage of the model, however, if it proves inadequate for characterizing item response data.

The Rasch model is a special case of Birnbaum's (1968) logistic model. Three types of logistic models might be distinguished according to the number of parameters. Birnbaum's three-parameter model assumes that the item characteristic curve can be specified in terms of a location parameter, an item discrimination parameter, and a parameter allowing for a non-zero lower asymptote. In the two parameter model, it is assumed that only the location and discrimination parameters are required, and in the Rasch model, it is assumed that only the location para- is required. Thus a natural question that needs to be addressed if the Rasch model were to be used for the problem of vertical equating is whether one

or both of the other parameters are necessary. Regardless of the number of parameters, all three logistic models assume that a unidimensional trait underlies the items.

Ignoring estimation problems, the three parameter logistic model is undoubtedly more adequate than the two parameter model or the Rasch model with only one parameter per item. Recent work by Lord (1975) suggests that in the long run the three-parameter logistic model may prove to provide a much improved means of vertical equating. The main disadvantages of the approach are the demands for very large sample sizes to achieve stable estimates and the considerable computing costs. The Rasch model is much simpler computationally than the three-parameter logistic model which would be a substantial advantage if the model provides an adequate approximation to real sets of data.

Following the notation of Wright and Panchapakesan (1969), the Rasch model specifies that the probability of a correct response to the i^{th} item by the n^{th} individual is

$$P_{ni} = Pr(a_{ni} = 1) = \frac{Z_n E_i}{1 + Z_n E_i},$$

where a_{ni} is the item score which takes a value of 1 if the response is correct and zero otherwise, Z_n is the ability score for the n^{th} person, and E_i is the item easiness. For most purposes, it is more convenient to deal with log ability ($b_n = \log Z_n$) and log easiness ($d_i = \log E_i$) which make it possible to express the log odds, L_{ni}, in the simple form

$$L_{ni} = \log \frac{P_{ni}}{1 - P_{ni}} = b_n + d_i.$$

As previously indicated, there are three assumptions of the Rasch model that may have questionable validity for typical multiple choice test items. That is, (1) the test may be multidimensional, (2) the items may vary in discriminating power, and (3) there may be a non-zero probability due to guessing of getting an item right regardless of the ability of the examinee. Wright (1968) acknowledged these three problems but argues that test construction should purposefully try to minimize them.

Some investigations of the robustness of the Rasch model under violations of the assumptions of equal discriminating power and lower asymptotes of zero have been conducted. Hambleton and Traub (1971) generated item response data based on the Birnbaum three-parameter logistic model. They then compared the results based on an assumed Rasch model and an assumed Birnbaum two-parameter model to those results based on the three parameters used to generate the data. Both the Rasch and the two-parameter Birnbaum models became noticeably less efficient when guessing was introduced. The two parameter model was generally more efficient than the Rasch model except at low ability levels under conditions of no guessing.

One of the potential advantages of the Rasch or other latent trait models over conventional equating procedures is the possibility that the item parameters and therefore the test calibration are invariant. That is, the estimates of the item parameters should not depend on the sample used to obtain the estimates which is what Wright (1967) refers to as "person-free test calibration." Several studies (e.g., Anderson, Kearney, and Everett, 1968; Tinsley and Dawis, 1975) have found that the Rasch item parameter estimates have relatively good invariance for particular sets of items. As might be expected, the invariance is improved when consideration is limited to those items that are found to fit the Rasch model within a given confidence interval.

Particularly relevant for the vertical equating problem are results such as those reported by Wright (1968) which compare estimates of ability based on "hard" and "easy" tests. This approach was used to investigate the adequacy of the "item-free person measurement" claim. Using test responses of 976 law students to a 48-item test, separate scores were obtained for each student based on the 24 easiest items and on the 24 hardest items. As would be expected, there was a substantial difference in the mean raw number right scores for the easy and hard tests (17.16 vs. 10.38 respectively). When estimated log ability scores were obtained, the means of the two tests were quite similar (means of 0.464 and 0.403 on the easy and hard tests respectively). To make a comparison between the difference in raw score means and the difference in log ability means, the differences in means can be compared to the corresponding standard deviations of the differences. For raw scores, the mean difference is 6.78 and the standard deviation of the difference is 3.30; thus, almost all the raw score differences are positive. For log ability on the other hand, the mean difference is 0.061, while the corresponding standard deviation is 0.749. The log ability differences are significantly greater than zero ($t = 2.54$) but the magnitude of the difference is small.

Whitely and Dawis (1974) report fairly similar results for a reanalysis of Tinsley's (1972) data for 949 subjects on 60 verbal analogy items. Again the items were divided into easy and hard subtests. The t for the difference in raw score on the easy and hard subtests was 42.52 whereas the corresponding value for the log ability scores was only 2.15.

Another comparison between "easy" and "hard" tests was made in both the Wright (1968) and Whitely and Dawis (1974) studies by converting scores to "standardized difference scores." The standard errors associated with a given individual's ability estimate on the hard and easy tests are used along with the two ability estimates to obtain a "standardized difference score", D_n, as follows:

$$D_n = \frac{b_{ne} - b_{nh}}{\sqrt{S^2_{ne} + S^2_{nh}}}$$

where b_{nh} and b_{ne} are the log ability estimates for individual n on the hard and easy tests respectively, and S^2_{nh} and S^2_{ne} are the estimated variances of the error of measurement associated with the individual's log ability estimate on the hard and easy tests. Wright and Panchapakesan (1969) provide an algorithm for obtaining the necessary estimated error variances in addition to the ability and item estimates of the Rasch model.

Using the D_n scores shown above, Wright (1968) computed means and standard deviations and noted that if the log ability estimates from the hard and easy tests were statistically equivalent, the mean should be zero and the standard deviation 1.0. The values actually obtained by Wright were 0.003 and 1.014 for the mean and standard deviation respectively. This result was judged to provide strong evidence for the equivalence of the hard and easy tests. Although the mean of 0.057 and the standard deviation of 1.146 reported by Whitely and Dawis (1974) are not as good as the values obtained by Wright, they do lend some support for the item-free person measurement claim of the Rasch model.

The results obtained by Wright and by Whitely and Dawis are very encouraging because of their potential significance for the vertical equating problem. There remain questions, however, about the generalizability of these results. It would be desirable to have more information about the consistency of the relative standing of a group of individuals on two equated tests that differ substantially in difficulty. It would also be desirable to have information about the stability of the results when estimates are obtained from one sample of examinees and then applied to a different sample of examinees. Finally, it would be helpful to have information on whether hard and easy tests are uniquely equated if divergent groups of examinees are used to perform the equating. Analyses of some existing item response data were undertaken in an attempt to provide just such information.

EMPIRICAL ANALYSES USING THE RASCH MODEL

Procedure

Item response data for 1,365 students on 50 items of a retired form of the College Entrance Board's Mathematics Achievement Test Level I were obtained from the files of the Office of Instructional Resources, Measurement and Research Division, of the University of Illinois.* This test was used as the intermediate mathematics proficiency and placement examination for all 1973 incoming freshmen at the University of Illinois who have not previously had a trigonometry course. Based on the 1,365 students, items 37-50 were discarded because of possible speededness or because the proportion of students correctly responding to a given item, p, was less than 0.20 or greater than 0.80. With but a few exceptions, items 37-50 had p values less than 0.15 and the ones which did not were very close to 0.20 and had associated proportions omitting equal to 0.40 or greater. Of the 36 items retained, the p values ranged from 0.22 to 0.77 except for two items which had p values of 0.82 and 0.81 with associated proportions omitting equal to 0.01 and 0.04 respectively. The p values were also used to create two subtests. An "easy" test consisted of the 18 items with the highest p values and a "difficult" test consisted of the 18 items with the lowest p values. In addition to eliminating several items, any student who responded correctly or incorrectly to all 36 items or to the two 18 item subtests was eliminated from all analyses. This was done because no information can be obtained for the item analyses from students who respond at these two extremes. Of the 58 students eliminated, it is, of course, possible that a student could have a score of 0 or 18 on one of the two subtests but be usefully included in the total test analyses, but for simplicity these few students were also eliminated.

The complete set of group/test combinations that were utilized in this study is summarized in Table 5-9. Nine sets of parameters were obtained corresponding to the crossing of the three possible tests (difficult, easy, and total) and the three examinee groups used for estimation (high, low, and total). As indicated in Table 5-9, these group/test combinations will be referred to by two letters identifying the test, then the group. For example, estimates based on the difficult test and the low group are labelled DL. The other possible labels are specified in Table 5-9.

The division of items into easy and difficult subtests is in line with the subtests used by Wright (1967) and one of the pairs of subtests investigated by Whitely and Dawis (1974) and therefore some of the analyses presented here parallel their analyses. However, in addition to using the total sample, three subpopulations of examinees were formed according to their "ability" level. The examinees were assigned to a "high" group if they had 21 or more items correct on the total 36-item test. With 16 or fewer items correct, examinees were assigned to a "low" group. The remaining examinees who had scores between 17 and 20 were retained in a "middle" group. This split assigned 490 examinees to the high group, 483 to the low group, and the remaining 334 to the middle group.

---

* We wish to thank Dr. David Frisbe for providing us with access to these data.

TABLE 5-9

Design for Rasch Estimates
of Item Parameters and Ability

	Group		
Test	High	Low	Total
Difficult	DH	DL	DT
Easy	EH	EL	ET
Total	TH	TL	TT

Item parameter estimates for all 36 items were obtained for each of the three groups via the Wright and Panchapakesan (1969) computer program. These three sets of 36-item parameter estimates were then used as the values of the item parameters for the easy and difficult tests for each of the three appropriate groups. For example, the 18-item parameter estimates corresponding to the 18 easiest items obtained for TT were used for ET and the other 18-item parameter estimates were used for DT. Ability estimates were then computed by the iterative Newton-Ralphson procedure given that the items were already calibrated. However in addition to obtaining ability estimates that used previously calibrated items, it was decided to compare these ability estimates with ones that used no prior information for the item parameters. The Pearson-product moment correlation between the two ability estimates was 1.0 for the total, high and low groups. Because of these three perfect correlations, only the results based on the ability estimates obtained by using the previously calibrated items are reported. The middle group was not used to obtain estimates (other than as part of the total group) but it was used to compare the equivalence of the easy and difficult tests by using the ability estimates based on the high group (and also the low group) and applying them to the middle group.

RESULTS

The results for the comparison of the difficult and easy tests for the total sample (DT and ET) are reported in Table 5-10. These results parallel those reported by Wright (1968) and by Whitely and Dawis (1974). As would be expected, the means on the two tests are quite different for the number right scores, but quite similar for the log ability scores. A t test for the difference in means on the number right score yields a value of 65.74, while the t for the difference in means on the estimated log ability score is only 1.82. Furthermore, the mean and standard deviation of the "standardized difference scores" are near 0.0 and 1.0 respectively as would be expected for statistically equivalent tests. Thus, based on the total sample the easy and difficult tests appear to be well equated on the log ability scale.

The above comparison of easy and difficult tests was repeated for both the high and low groups. These results are reported in Table 5-11. The results for the estimated log ability scores for the high and low groups are less favorable than those for the total group but they still provide reasonably good support for the claim that the scale provides equivalent measurement. The main exceptions to the support for equivalent measurements come from two sources: (1) the relatively large mean of the standardized difference scores obtained for the high group, and (2) the relatively large discrepancy between 1.0 and the standard deviations of 0.932 and 1.115 obtained for the standardized difference scores for the high and low groups respectively.

One of the requirements stated early in this section for equating is that the conversion from raw to scale scores be unique for different subpopulations. To investigate this assertion, the independent conversions for the high and low groups were compared. If the log ability estimate associated with a particular number right score for the high group

TABLE 5-10

Comparison of Difficult and Easy Test Results for the Total Group

Statistic	Easy Test	Difficult Test	Difference	Standardized Difference
		Number Right Score		
Mean	11.975	6.514	5.461	
Std. Error	0.098	0.086	0.083	
Std. Dev.	3.539	3.127	3.003	
		Estimated Log Ability		
Mean	0.114	0.069	0.045	-0.023
Std. Error	0.030	0.025	0.025	0.029
Std. Dev.	1.090	0.903	0.897	1.039

TABLE 5-11

Comparison of Difficult and Easy
Test Results for the High and Low Groups

Statistic	Easy Test	Difficult Test	Difference	Standardized Difference
		Number Right Score (High Group)		
Mean	15.131	9.500	5.631	
Std. Error	0.063	0.108	0.119	
Std. Dev.	1.402	2.387	2.634	
		Number Right Score (Low Group)		
Mean	8.453	3.797	4.656	
Std. Error	0.126	0.072	0.149	
Std. Dev.	2.773	1.591	3.271	
		Estimated Log Ability (High Group)		
Mean	0.995	1.005	-0.010	-0.093
Std. Error	0.030	0.028	0.037	0.042
Std. Dev.	0.662	0.611	0.828	0.932
		Estimated Log Ability (Low Group)		
Mean	-0.809	-0.832	0.023	-0.037
Std. Error	0.034	0.029	0.045	0.051
Std. Dev.	0.745	0.631	0.978	1.115

is plotted against the estimate for the low group, the points should fall on a straight line through the origin with a slope of one if the conversion is unique. The results of such a plotting of ability estimates are given in Figures 5-1 and 5-2 for the easy and difficult tests respectively.

Inspection of Figures 5-1 and 5-2 shows that with the notable exception of the lowest scores on the easy test (Figure 5-1), the points fall very nearly on a 45° line through the origin. By far the largest exception is for the lowest raw score on the easy test (Figure 5-1) where the estimated log ability based on the high group is much too low compared to the estimated log ability based on the low group. This exception occurs at the lowest score on the easy test where the standard error of estimate for the high group is very large. Thus, the exception may not be considered very serious. In general, the results in Figures 5-1 and 5-2 are in close agreement with the results previously reported by Anderson, et al (1968) and by Tinsley and Dawis (1975).

The uniqueness of the equating of easy and difficult tests for different groups may be evaluated more directly by comparing the equating lines obtained for different groups. A linear equating of the estimated log ability estimates based on the easy and difficult tests yields the solid line shown in Figure 5-3 for the high group and the dashed line for the low group. These two lines would coincide if the same conversion applied to both groups. While the lines in Figure 5-3 are reasonably close, there are noticeable differences at the high ability levels. For example, an estimated log ability of 2.0 on the difficult test would be linearly equated to an estimated log ability of about 2.1 on the easy test when the equating is based on the high group (solid line). The comparable values when linear equating is based on the low group, however, are 2.0 and 2.5. The reason for this discrepancy can be seen by referring to the values of the standard deviations reported in Table 5-11. As noted in Table 5-11, the standard deviations of the log ability scores are more discrepant from easy to difficult tests for the low group than are the corresponding standard deviations for the high group.

The results discussed so far suggest that the Rasch model provides at least a rough equating of the two subtests which differ markedly in difficulty level. Since the subtests differ more in difficulty than would adjacent levels of a test to be vertically equated, it might still be argued that the approach has potential value for the vertical equating problem. It should be recalled, however, that while the easy and difficult tests may be roughly equivalent statistically, they differ substantially in their precision for the different levels of ability.

As a final comparison of the Rasch results for tests of different difficulty and groups of different ability, the parameter estimates obtained for high and low groups were applied to the examinees in the middle group. This provides an evaluation of the adequacy of the equating of tests of different difficulty when the estimates obtained from one group are applied to a group at an adjacent ability level.

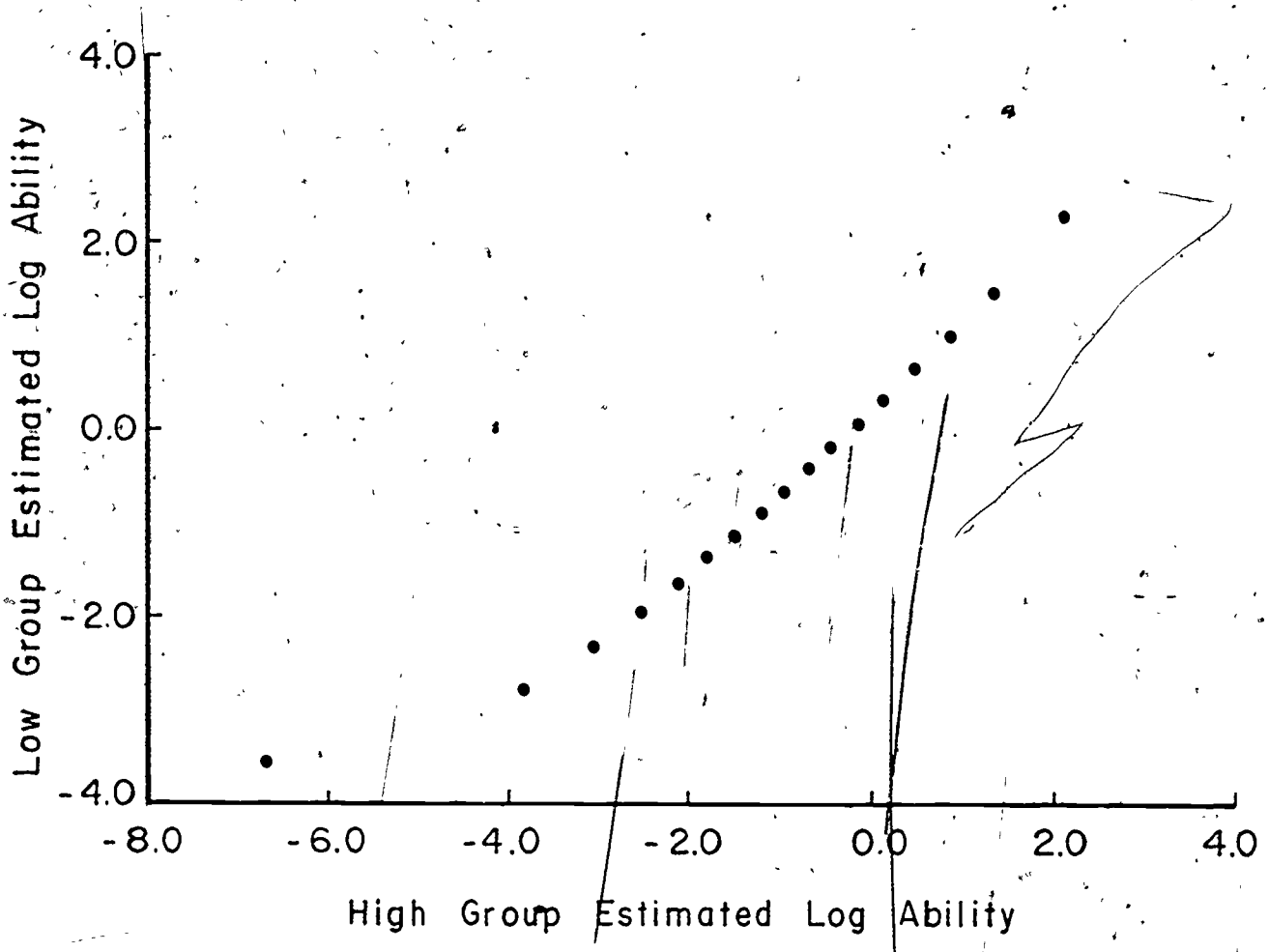


Figure 5-1

Plot of Easy Test Conversions

to Estimated Log Ability

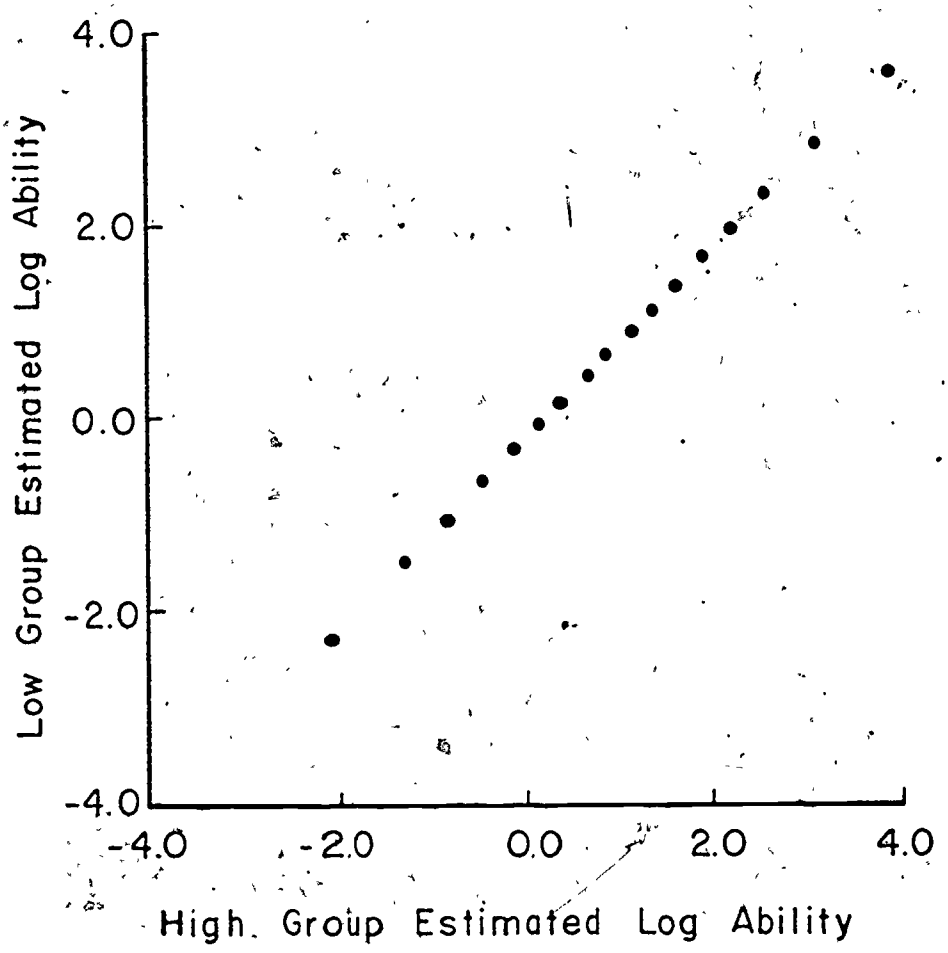


Figure 5-2

Plot of Difficult Test Conversions

to Estimated Log Ability

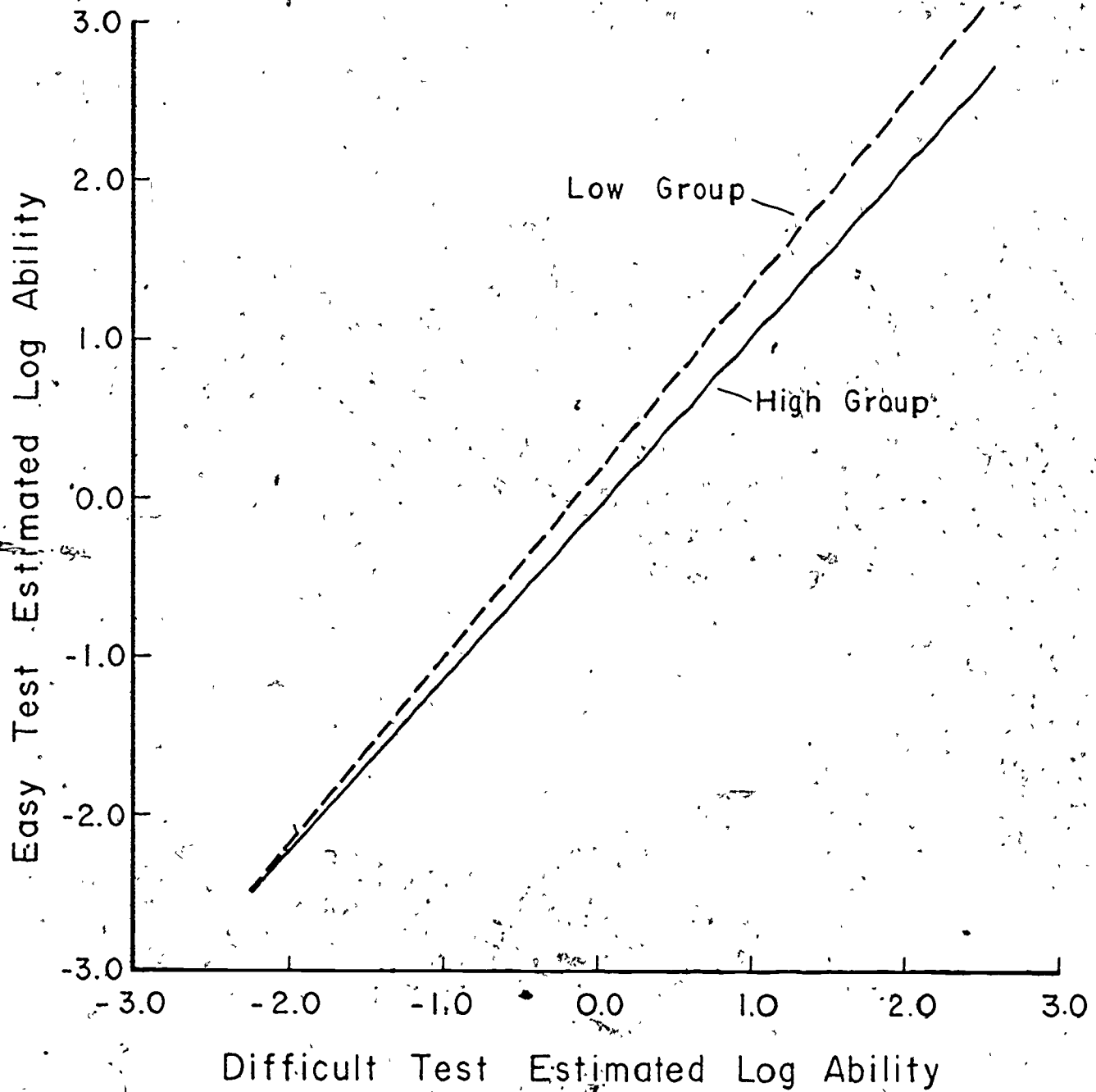


Figure 5-3

Equated Estimated Log Ability Lines

for High and Low Groups

The means, standard errors, and standard deviations for the middle group on the easy and difficult tests are reported in Table 5-12. The three sections of Table 5-12 provide the results for number right scores, estimated log ability based on high group data, and estimated log ability based on low group data. As was the case earlier, the means on the two tests are quite different for the number right scores ($t = 41.51$). However, the results based on the log ability estimates are not as good as the corresponding results reported when ability estimates were applied to the same group. The value of t for the difference between means on the easy and difficult tests is -3.38 when the ability estimates obtained from the high group were applied to the middle group. When the ability estimates obtained from the low group were applied to the middle group, $t = 7.34$. The magnitude of these differences between means is not trivial which leads to the following generalization. A middle group examinee would do better to take the hard test when ability estimates are obtained from the high group, but would do better to take the easy test when the estimates are obtained from the low group. This is not a very desirable feature for two tests that are to be vertically equated. In addition, even though the standard deviations of the standardized difference scores are near 1.0 when either type of ability estimates are used, the means do differ significantly from 0.0 in both cases. Clearly, the two tests cannot be regarded as statistically equivalent. Therefore, based on the results of obtaining ability estimates from one group and applying these same estimates to a different group, the easy and difficult tests do not seem to provide equivalent measurements which are so necessary for longitudinal research.

CONCLUSIONS

Based on a logical analysis as well as the empirical comparisons of scaled scores on different levels of standardized tests, which according to the results of the Anchor Test Study have "equivalent" raw scores, it must be concluded that the vertical equating of existing tests is often less than satisfactory. Lord (1975) has suggested that among current methods of equating, only those based on item characteristic curve theory (i.e., latent trait models) are appropriate for the task of vertical equating. Of these, the Rasch model is probably the simplest. But, our empirical results raise doubts about the adequacy of this model, at least, for some sets of test items.

The empirical analyses involving the Rasch model that are presented above do not support the dual claims of item-free person measurement and person-free test calibration. It may be that the comparisons reported above were more extreme, in terms of the wide separation of the high and low groups than are apt to be encountered when equating tests over adjacent grades. Also, better results might be expected by use of an anchor test procedure. Thus, the test may be overly severe. It is also possible that more careful selection of items that fit the model is necessary, which is the approach that seems to be suggested by Keats and Boldt as reported by Angoff (1971, pp. 529-530).

More work on the vertical equating problem using latent trait models is clearly needed. This should include tests of the limits of

TABLE 5-12

Comparison of Difficult and Easy Test Results for the Middle Group

Statistic	Easy Test	Difficult Test	Difference	Standardized Difference
		Number Right Score		
Mean	12.437	6.063	6.374	
Std. Error	0.082	0.083	0.154	
Std. Dev.	1.495	1.510	2.806	
		Estimated Log Ability Based on High Group Data		
Mean	-0.029	0.124	-0.154	-0.276
Std. Error	0.026	0.023	0.046	0.057
Std. Dev.	0.474	0.420	0.835	1.040
		Estimated Log Ability Based on Low Group Data		
Mean	0.233	-0.090	0.323	0.356
Std. Error	0.023	0.023	0.044	0.054
Std. Dev.	0.448	0.415	0.804	0.978

applicability of the Rasch model as well as investigations of models involving more parameters. Additional work involving overlapping groups and the use of an anchor test approach is currently underway.

REFERENCES

Anderson, J., Kearney, G. E., & Everett, A. V. An evaluation of Rasch's structural model for test items. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (ed.), Educational Measurement, 2nd Edition, Washington, D. C.: American Council on Education, 1971.

Bianchini, J. C. & Loret, P. G. Anchor Test Study. Final Report. Project Report and Volumes 1 through 33. ERIC Documents ED 092 601 through ED 092 634, 1974.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Norvick, Statistical Theories of Mental Test Scores, Reading, Massachusetts: Addison-Wesley, 1968, chapters 17-20.

CTB/McGraw-Hill, California Achievement Tests, 1970 edition, Monterey, California: CTB/McGraw-Hill, 1970.

CTB/McGraw-Hill, Comprehensive Tests of Basic Skills, 1968 edition, Monterey, California: CTB/McGraw-Hill, 1968.

Educational Testing Service. School and College Ability Test, Princeton, New Jersey: Educational Testing Service, 1957.

Hambleton, R. K. & Traub, R. E. Information-curves and efficiency of three logistic test models. British Journal of Mathematical and Statistical Psychology, 1971, 24, 273-281.

Harcourt, Brace, Jovanovich. Metropolitan Achievement Tests, 1970 edition, New York: Harcourt, Brace, Jovanovich, 1970.

Harcourt, Brace, Jovanovich. Stanford Achievement Tests, 1973 edition, New York: Harcourt, Brace, Jovanovich, 1973.

Lord, F. M. A survey of equating methods based on item characteristic curve theory (ETS RB 75-13). Princeton, New Jersey: Educational Testing Service, 1975.

McCarthy, P. J. Replication: An Approach to the Analysis of Data from Complex Surveys, Washington, D. C.: National Center for Health Statistics, Vital and Health Statistics, Series 2, No. 14, 1966.

Rasch, G. Probalistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research, 1960.

Rasch, G. An individualistic approach to item analysis. In P. F. Lazarsfeld and N.W. Henry (eds.), Readings in Mathematical Social Science, Chicago: Science Research Associates, 1966a, pp. 89-108.

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 1966b, 19, 49-57.

Tinsley, H. E. An investigation of the Rasch simple logistic model: sample free item and test calibration. Educational and Psychological Measurement, 1975, 35, 325-339.

Whitely, S. E. & Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Wright, B. D. Sample free test calibration and person measurement. Proceedings of the 1967 Invitational Conference on Testing Problems, Princeton, New Jersey: Educational Testing Service, 1968, pp. 85-101.

Wright, B. D. & Panchapakesan, N. A procedure for sample free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

CHAPTER 6

APPLICATIONS OF THE SIMPLEX MODEL IN LONGITUDINAL STUDIES

In a variety of situations where repeated measurements are obtained over several points in time, the intercorrelation matrix has been observed to have particular characteristics. Typically the correlations between measures obtained at adjacent points in time are found to be higher than the correlations between measures that are further apart in time. This pattern of correlations is also characteristic of Guttman's simplex (1955) and a number of authors have suggested that the simplex is a good model for explaining change over time (e.g., Humphreys, 1960, 1968; Jones, 1962).

One of the difficulties that investigators have had in evaluating the adequacy of the simplex model for a set of correlational data is that the correlations are attenuated due to errors of measurement. While the simplex model may be appropriate for error free measures, the fit to correlations of fallible measures may be poor due to the errors of measurement. Humphreys (1960) recognized this problem and tried to deal with it by estimating reliability coefficients.

Another difficulty in evaluating the fit of a simplex model to a set of empirical data is, of course, sampling error. Jöreskog (1970) developed estimation techniques for a variety of simplex models including the model most commonly postulated for growth data which he refers to as a quasi-Markov simplex. For example, the quasi-Markov simplex corresponds to the one suggested by Humphreys (1960). Jöreskog's estimation procedures (e.g., Jöreskog, Gruvaeus, and van Thillo, 1970; Jöreskog and van Thillo, 1972) provide maximum likelihood estimates which allow for errors of measurement and yield large sample chi square tests based on an assumption of multivariate normality.

Recently Werts, Linn and Jöreskog (in press, a) have shown that the simplex model provided a reasonably good fit to the intercorrelations of achievement test results reported by Bracht and Hopkins (1972). Those data were obtained on a yearly basis over grades 1 through 9. Werts, Linn and Jöreskog (in press, b) have also used the simplex model to analyze the intercorrelations of grades in college over 8 semesters that were reported by Humphreys' (1968). This reanalysis confirmed Humphreys' assertion that the data fit a simplex model. Humphreys' belief that the reliabilities of grades across semesters were equal was also supported by the analyses.

In this chapter the simplex model will be briefly reviewed within the context of longitudinal studies. Procedures for estimating model parameters as well as correlations of gain with status at an earlier point in time will be discussed. Finally, the results of application of the simplex model to several sets of longitudinal data will be reported.

THE MODEL

The simplex model can be represented in several ways (see for example Corballis, 1965, Jöreskog, 1970). A conceptually appealing form for growth data, however, is to assume that, in the absence of errors of measurement, a score at time t + 1 is a function of the score at time t plus an uncorrelated increment. More specifically, a person's true score at time t + 1, Z_{t+1}, is assumed to be

$$Z_{t+1} = b_t Z_t + U_{t+1} \qquad (6.1)$$

where U_{t+1} is assumed to be uncorrelated with Z_t. If, for convenience, the Z's are all standardized, then the correlation between Z_t and Z_{t+1} is simply b_t and the correlation between Z_i and Z_j $(i > j)$ is the product of the b_t for $t = j, j+1, \ldots, i$.

It should be noted that the assumption that U_{t+1} and Z_t are uncorrelated does not imply that growth is uncorrelated with previous status as has sometimes been assumed. Still dealing with the error free measures, Z_t, the usual definition of growth from time t to time t + 1 is,

$$Z_{t+1} = Z_t + \Delta_{t+1}, \qquad (6.1)$$

where Δ_{t+1} is the change or "growth". The change in equation 6.2, Δ_{t+1}, can be expressed in terms of the components of equation 6.1 as follows:

$$\Delta_{t+1} = (b_t - 1) Z_t + U_{t+1}. \qquad (6.3)$$

From equation 6.3 the covariance of Δ_{t+1} and Z_t is readily obtained

$$\sigma(Z_t, \Delta_{t+1}) = (b_t - 1)\, \sigma^2(Z_t), \qquad (6.4)$$

where $\sigma^2(Z_t)$ is the variance of Z_t.

From equation (6.4) it is clear that the correlation between status at time t and growth will be zero only when $b_t = 1$. Typically, b_t will not equal 1.0, hence the correlation between status at time t and growth will be non-zero.

The fallible observed measures are assumed to follow a classical test theory model at any point in time. Thus, an observed score, X,

at time t may be represented by.

$$X_t = Z_t + e_t \qquad (6.5)$$

where e_t is assumed to have an expected value of zero and to be uncorrelated with Z at all points in time and uncorrelated with e at points in time other than t. The model may be depicted by a path analysis diagram as shown in Figure 6-1.

The usual observed gain score, D_{t+1}, is simply the difference between the observed score at time t + 1 and the observed score at time t. Thus,

$$D_{t+1} = X_{t+1} - X_t , \qquad (6.6)$$

which in terms of the true change, Δ_{t+1}, and errors of measurement is

$$D_{t+1} = \Delta_{t+1} + (e_{t+1} - e_t) . \qquad (6.7)$$

Equations (6.6) and (6.7) are the standard equations for a simple gain score expressed respectively in terms of observed scores and in terms of true gain and errors of measurement. As such, equations (6.6) and (6.7) are independent of the assumed underlying simplex model on the error free measures. The relationship of D_{t+1} to the parameters of the simplex model can be seen by substituting equation (6.3) into equation (6.7).

MATRIX FORMULATION

The model as outlined above implies a particular structure for the observed score variances and covariances. This structure is most conveniently represented in matrix form (see for example Joreskog, 1970; Werts, Linn & Joreskog, in press, a). Let,

$$\underset{\sim}{X}' = [X_1, X_2, \ldots X_p]$$

be in row vector of observed scores at p points in time,

$$\underset{\sim}{e}' = [e_1, e_2, \ldots, e_p]$$

be a row vector of errors of measurement, and

$$\underset{\sim}{Z}' = [Z_1, Z_2, \ldots, Z_p]$$

be a row vector of true scores. The vector of observed scores is simply

$$\underset{\sim}{X} = \underset{\sim}{Z} + \underset{\sim}{e} . \qquad (6.8)$$

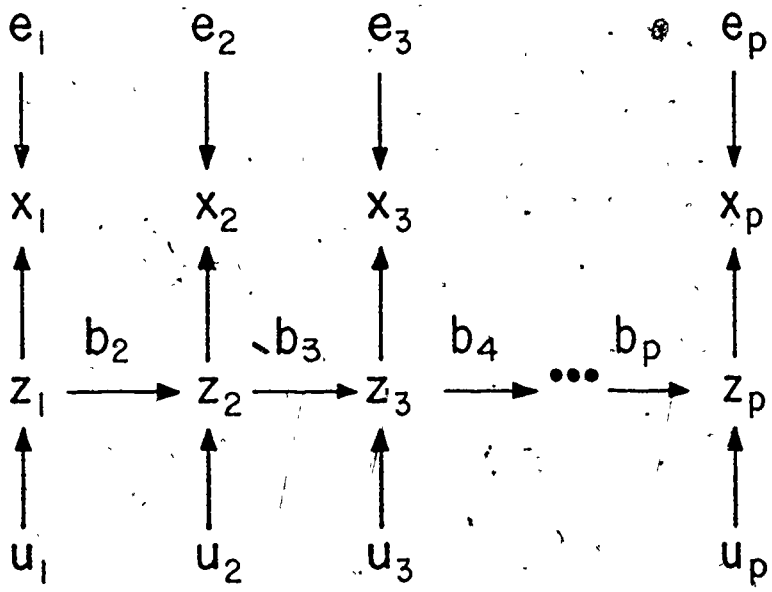


Figure 6-1

Path Analysis Diagram of Quasi-Simplex Model

In order to relate the observed scores to the parameters of the simplex model let

$$\underset{\sim}{U}' = [U_1, U_2, \ldots U_p]$$

be a row vector of the uncorrelated increments and let B be a p x p matrix, with unities down the main diagonal, with elements $-b_1$, $-b_2$, $\ldots -b_{p-1}$ next to the main diagonal on the lower left hand side, and zeros elsewhere. For example, with p = 5, the B matrix is

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ -b_1 & 1 & 0 & 0 & 0 \\ 0 & -b_2 & 1 & 0 & 0 \\ 0 & 0 & -b_3 & 1 & 0 \\ 0 & 0 & 0 & -b_4 & 1 \end{bmatrix}$$

With these definitions and equation (1) the relationship of Z and U is given by

$$B\underset{\sim}{Z} = \underset{\sim}{U}$$

assuming $\underset{\sim}{Z}_0 = 0$. The simplex model on the error free parameters can now be written as

$$\underset{\sim}{Z} = B^{-1}\underset{\sim}{U} \qquad (6.9)$$

Since B^{-1} is a lower triangular matrix with entries as illustrated below for the case of p = 5;

$$B^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ b_1 & 1 & 0 & 0 & 0 \\ b_1b_2 & b_2 & 1 & 0 & 0 \\ b_1b_2b_3 & b_2b_3 & b_3 & 1 & 0 \\ b_1b_2b_3b_4 & b_2b_3b_4 & b_3b_4 & b_4 & 1 \end{bmatrix}$$

it can be seen that this formulation is equivalent to setting $Z_1 = U_1$ but except for this additional specificity equations (6.1) and (6.9) are equivalent.

The variance covariance matrix among the p observed variables, Σ, can now be specified in terms of the parameters of the simplex model and the variances of the errors of measurement,

$$\Sigma = B^{-1} \Psi B'^{-1} + \Theta^2 , \qquad (6.10)$$

where Ψ is a diagonal matrix with the variances of the U_t as entries, $\sigma^2(U_t)$, and Θ^2 is a diagonal matrix with variances of the errors of estimate as entries, $\sigma^2(e_t)$.

ESTIMATES

Estimates of matrices involved in (6.10) will be denoted by a hat over the corresponding population matrix in (6.10) Thus,

$$\hat{\Sigma} = \hat{B}^{-1} \hat{\Psi} \hat{B}'^{-1} + \hat{\Theta}^2 . \qquad (6.11)$$

Unfortunately, several of the elements of the three matrices on the right hand side of (6.11) are not identified (Jöreskog, 1970). To achieve identification some additional restrictions are required. One possibility is to arbitrarily assign fixed values to $\sigma^2(e_1)$ and $\sigma^2(e_p)$. When this was done by Werts, Linn, and Jöreskog (in press, a) the parameter estimates for the remaining elements provided a good fit to the observed variance covariance matrix.

An alternative approach to obtaining unique estimates is to add a restriction to the model that the variances of the errors of measurement are constant over time. That is, it is assumed that $\sigma^2(e_t)$ equals $\sigma^2(e)$ for all t. With this assumption, maximum likelihood estimates of the b_t, the $\sigma^2(u_t)$ and of $\sigma^2(e)$ may be obtained using the ACOVS program (Jöreskog, Gruvaeus, and van Thillo, 1970). Also obtained is a chi-square test of the model based on an assumption of multivariate normality. With this formulation there are $p(p + 1)/2$ unique elements in $\hat{\Sigma}$ and $2p$ parameters to be estimated (i.e., $p - 1$ values of the b_t, p values of $\sigma^2(U_t)$, and one value of $\sigma^2(e)$). This leaves $(p^2 - 3p)/2$ degrees of freedom for the chi-square test.

With large samples, the chi-square test will often be of less interest than the magnitude of the discrepancies between the variance-covariance matrix implied by the parameter estimates of the model, $\hat{\Sigma}$, and the observed sample variance-covariance matrix, S. With variables that have arbitrary variances as is frequently the case in the social sciences the sam-

ple correlation matrix, R, and the corresponding matrix implied by the model, $\hat{R}$, will often be of greater interest. The residual matrix is simply the difference between the observed correlation matrix, R, and the estimate of the observed correlation matrix, $\hat{R}$, that is implied by the model parameter estimates. With large sample sizes the residual matrix is of special interest since the chi-square test will typically lead to a rejection of the model. A significant chi-square is to be expected for any a priori model such as the above given a sufficiently large sample size. For evaluating the adequacy of the model it is important also to consider the magnitude of the deviations from the model. The residual matrix provides this information. If a single index of fit is desired, the root mean square of the residuals is sometimes useful (see for example, Linn and Werts, in press).

GROWTH STATISTICS

If it is decided that the fit of the data to the model is adequate, the parameter estimates may be used to estimate a variety of statistics that are ordinarily considered to be of interest in longitudinal studies. For example, the estimated correlation between true change from time t t to time t + 1 with status at time t is

$$\hat{\rho}(\Delta_{t+1}, Z_t) = (b_t - 1) \frac{\hat{\sigma}(Z_t)}{\hat{\sigma}(\Delta_{t+1})}, \qquad (6.12)$$

where $\hat{\sigma}(\Delta_{t+1})$ is the estimated standard deviation of true change which is given by

$$\hat{\sigma}(\Delta_{t+1}) = \sqrt{\hat{\sigma}^2(Z_t) + \hat{\sigma}^2(Z_{t+1}) - 2\hat{\sigma}(Z_t Z_{t+1})}.$$

The estimated reliability of the simple gain scores is

$$\hat{\rho}^2(\Delta_t, D_t) = \frac{\hat{\sigma}^2(\Delta_t)}{\hat{\sigma}^2(\Delta_t) + \sigma^2(e_t) + \sigma^2(e_{t-1})} \qquad (6.13)$$

A potential advantage of formulas such as (6.12) and (6.13) over the traditional estimates is that they are based on all data points rather than just two points in time. This is only an advantage, however, to the degree that the model is adequate for the data.

Estimated covariances or correlations between true status at any two points in time, say t and t + k, may be obtained from the model parameter estimates as follows:

$$\hat{\sigma}(Z_{t+k}, Z_t) = \hat{b}_{t+k-1}\,\hat{b}_{t+k-2} \cdots \hat{b}_t \hat{\sigma}^2(Z_t)$$

and

$$\hat{\rho}(Z_{t+k}, Z_t) = \hat{b}_{t+k-1}\,\hat{b}_{t+k-2} \cdots \hat{b}_t\, \hat{\sigma}(Z_t)/\hat{\sigma}(Z_{t+k}).$$

The covariance of Z_{t+k} and Z_t along with the variance of Z_t can in turn be used to estimate the covariance of the true change from time t to t + k, Δ_{t+k}, and initial status:

$$\hat{\sigma}(\Delta_{t+k}, Z_t) = \hat{\sigma}(Z_{t+k}, Z_t) - \hat{\sigma}^2(Z_t).$$

If there were no errors of measurement the measures at time t and time t + 1 would contain all the information about Δ_{t+1}. With errors of measurement, however, the observed scores at times other than t and t + 1 may contribute to the prediction of Δ_{t+1}. Thus, if there were an interest in obtaining estimated true gains between t and t + 1 then all the observed scores $X_1, X_2, \ldots X_p$ might be used as predictors as is implied by Cronbach and Furby (1970) and Werts, Joreskog and Linn (1972). Estimated covariances of observed scores with the true change may be obtained using the model parameters. These covariances along with the observed score variance-covariance matrix could then be used to obtain multiple regression estimates of Δ_{t+1}. The resulting estimate would have to be at least as good as the more natural estimate obtained from X_t and X_{t+1} alone. This result is of little comfort, however, because, as shown by Tatsuoka (1975), the multiple regression estimate of Δ_{t+1} based $X_1, X_2, \ldots X_p$ will be better than the one based on X_t and X_{t+1} only if the errors of measurement are correlated, which, of course, violates the assumptions of the model.

Table 6-1.

High School Rank and Grade Point Averages for Eight Semesters of College

a. Intercorrelations

Semester	HS	1	2	3	4	5	6	7	8
HS	1.000								
1	.387	1.000							
2	.341	.556	1.000						
3	.278	.456	.490	1.000					
4	.270	.439	.445	.562	1.000				
5	.240	.399	.418	.496	.512	1.000			
6	.256	.415	.383	.456	.469	.551	1.000		
7	.240	.387	.364	.445	.442	.500	.544	1.000	
8	.222	.342	.339	.345	.416	.453	.482	.541	1.000

b. Residuals ($R-\hat{R}$)

Semester	HS	1	2	3	4	5	6	7	8
HS	.000								
1	.013	-.009							
2	-.010	.001	.004						
3	-.016	-.008	.007	.004					
4	-.012	-.006	-.018	.007	.009				
5	-.012	.001	.004	-.001	-.001	.001			
6	.019	.041	-.006	-.010	-.013	.005	.000		
7	.021	.041	.003	.013	-.005	-.006	.004	-.009	
8	.024	.029	.013	-.045	.013	-.004	-.005	.006	.000

Table 6-1 (Continued)

c. Parameter Estimates

Semester	Beta	Var(u)
HS	--	.583
1	.642	.350
2	.939	.057
3	.836	.175
4	.958	.041
5	.894	.122
6	.940	.070
7	.926	.092
8	.904	.100

var(e) = .417

Chi-Square = 40.07 with 27 d.f. (p = .051)

EXAMPLE OF FIT (ACADEMIC ACHIEVEMENT)

Humphreys (1968) observed that the intercorrelations of high school grades and grades in eight semesters of college followed a pattern typical of a simplex. Werts, Linn, and Jöreskog (in press, b) reanalyzed Humphreys data using a simplex model and found a good fit. For illustrative purposes another analysis of these data, which are based on a sample of approximately 1,600 students is reported below. The model differs slightly from that used by Werts, Joreskog and Linn.

The specific model used with these data is the same as equation (6.11) except that the procedure used the sample correlation matrix rather than a variance-covariance matrix. The restriction that the variances of the errors of estimate are equal was used. A total of 9 variables (high school grades plus 8 semesters of college grades) were used in the analysis.

The observed correlation matrix, R, is reported in section "a" of Table 6-1. As can be seen there is a clear tendency for the correlations among adjacent semesters (entries next to the main diagonal) to be higher than the correlations between grades in more distant semesters. There are some reversals in the pattern, but generally the correlations get smaller as you move down a column, from right to left in a row, or from the main diagonal to the lower left hand corner of the triangular section of the correlation matrix shown in Table 1a.

Based on the observation of the correlation pattern a reasonably good fit to the simplex model might be expected. That this is the case is supported by the chi-square value of 40.07 which with 27 degrees of freedom has an associated p of approximately .051. While almost significant at the .05 level, with such a large sample size this would appear to be a quite good fit. Further support for the goodness of fit can be obtained from an inspection of section "b" of Table 2 which lists the residual elements (i.e., $R-\hat{R}$). None of the 45 residuals in Table 1b exceed .05 in absolute value and the root mean square of the residuals is only .015. Thus, these data fit the simplex model quite well even with the added assumption that the error variances are equal at all nine observation points.

The estimated correlations between true status at time t and true change from time t to time t + 1 are reported in Table 6-2. Also reported in Table 6-2 are the estimated reliabilities of the observed difference scores for each time interval. All of the correlations of true status with true change are negative. It should be noted, however, that this result is a consequence of two features of this particular analysis: (1) using standardized observed scores (i.e., a correlation rather than a variance-covariance matrix) and (2) restricting the error variances to be equal. Under these conditions the estimated variances of the true scores will be nearly equal and the value of b_t will be less than 1.0 which yields a negative correlation between true status and true change (see equation 6.12).

Table 6-2

Estimated Correlations Between True Change with Previous Status and Reliability of Change (Grade Data)

Time Interval of Change	Correlation of Δ_{t+1} with Z_t	Reliability of Change
1 to 2	- .42	.34
2 to 3	- .19	.07
3 to 4	- .29	.19
4 to 5	- .16	.05
5 to 6	- .22	.13
6 to 7	- .17	.08
7 to 8	- .18	.10
8 to 9	- .23	.11

The reliability of the change scores reported in Table 6-2 are all quite low. As would be expected, the reliability of the change is highest for time 1 (High School Rank) to time 2 (first semester college grades) which has the lowest correlation between adjacent times. The saw tooth pattern of the reliabilities for changes from adjacent semesters in college is relatively consistent with the pattern of same versus different academic years for the adjacent semesters. The reliability of change from one semester to another tends to be slightly lower if the two semesters are in the same academic year than if they involve two academic years. This corresponds to a tendency for grades in adjacent semesters in a single academic year to correlate somewhat higher than those involving different academic years. The most notable feature of these reliabilities, however, is their extremely low magnitude.

Another set of academic achievement data that illustrate the use of the simplex model were originally reported by Bracht and Hopkins (1972). Their data consisted of achievement test scores obtained at eight points in time (grades 1, 2, 3, 4, 5, 6, 7, and 9). The scores were reported in grade equivalent units. Thus, the scores at least have the superficial appearance of a common scale.

A previous attempt to fit these data to a simplex model (Werts, Jöreskog, and Linn, in press, a) resulted in a significant chi-square with $p = .035$. Due to the relatively large sample size (over 300) the significant chi-square is probably of less interest than the magnitude of the residuals. Based on the residuals and the root mean square of the residuals, however, the fit was judged to be reasonably good.

Since the detailed analysis of the Bracht and Hopkins data will be reported elsewhere (Werts, Linn, & Jöreskog, in press, a), they will not be repeated here. One aspect of the results that stands in sharp contrast to the above results for college grades is worthy of special note, however. The correlations of true status with true gain and the reliabilities of the gains were quite different in the Bracht and Hopkins data than they were in Humphreys' grade data. These correlations and reliabilities are reported in Table 6-3. As can be seen in Table 6-3, the correlations between true status and true gain are positive in all cases which contrasts with the negative correlations reported in Table 6-2. Also, the reliabilities of the difference scores reported in Table 6-3 are higher than the ones reported in Table 6-2.

As previously noted, the negative correlations of status and change reported in Table 6-2 are a result of analyzing correlations rather than covariances and of restrictions of the model. Since the variance-covariance matrix was analyzed for the results in Table 6-3 the estimated correlations might be either positive or negative. The fact that they are all positive is a result of a particular property of the grade equivalent scale which was discussed in Chapter 4 in this report. That is, the variance of the grade equivalent scale

Table 6-3

Estimated Correlations Between True Change with Previous Status and Reliability of Change (Bracht and Hopkins data)

Time Interval of Change	Correlation of Δ_{t+1} with Z_t	Reliability of Change
2 to 3	.67	.42
3 to 4	.12	.56
4 to 5	.59	.39
5 to 6	.09	.51
6 to 7	.22	.43

increases with grade level. This increase in variance with grade level not only results in positive correlations between observed initial status and observed change but between true initial status and true change. Whether substantive meaning should be attached to these positive correlations depends on one's view of the meaningfulness of increased variance with grade level.

The higher reliabilities of the change scores in Table 6-3 than in Table 6-2 are primarily due to the higher reliabilities of the achievement tests than of the grades. The achievement test reliabilities are in the 80's and 90's whereas the assumed common reliability of grades is estimated to be only .58.

At least for the two examples mentioned above, the simplex model appears to yield estimates that fit the observed data reasonably well. When this is true, the model has the advantage of requiring only a single measure of a construct at each point in time. Alternative models which are considered elsewhere in this report generally require multiple measures at each point in time. As will be seen below, the simplex model, at least in the simple form used to analyze the data proves to be relatively good for some sets of longitudinal data but relatively poor for others.

ABILITY MEASURES

Although the distinction between aptitude and achievement is one more of degree than of kind, it remains of interest to test the fit of the simplex model for tests that are closer to the basic aptitude end of the continuum than the achievement end. Aptitude tests may be distinguished from achievement tests primarily in terms of breadth of relevant experience and recency of learning with measures at the achievement end of the continuum being narrower and more recent (Humphreys, 1973). There is no good basis for postulating that aptitude is fixed. Indeed, as implied by Anderson (1939) and more formally specified by Humphreys (1960), there is reason to believe that the simplex model might be quite appropriate for aptitude measures. An attempt was made to fit two sets of data involving ability measures at the aptitude end of the continuum to the simplex model. The matrices of intercorrelations for both sets of data were obtained from Humphreys (1967).

The first set of data involves vocabulary test scores for 278 children obtained yearly from grades 2 through 6. The incorrelations among the vocabulary scores over these five points in time are reported in section "a" of Table 6-4. Inspection of the correlation matrix suggests that the simplex model may not be very adequate for these data. This is suggested by a number of instances where the correlation between scores obtained for grades separated by more time are as high or higher than those obtained for grades that are separated by less time.

Table 6-4

Vocabulary Scores from Grade to Grade

(N = 278)

a. Intercorrelations

Grade	2	3	4	5	6
2	1.00				
3	.65	1.00			
4	.58	.65	1.00		
5	.63	.73	.72	1.00	
6	.56	.68	.65	.76	1.00

b. Residuals $(R-\hat{R})$

Grade	2	3	4	5	6
2	.000				
3	.005	-.005			
4	-.010	-.027	.044		
5	.018	.027	.001	-.039	
6	-.021	.013	-.032	.024	.000

c. Parameter Estimates

Grade	Beta	Var(u)
2	--	.737
3	.876	.176
4	.913	.074
5	1.039	.030
6	.948	.038

Var(e) = .263

Chi Square = 17.76 with 5df (p = .003)

The parameter estimates for the simplex model are reported in section "c" of Table 6-4 along with the chi-square test. The residuals (i.e., $R-\hat{R}$) are reported in section "b" of Table 4. The chi-square value is significant at the .01 level which suggests that the model may not be adequate for these data. Given the relatively large sample size, however, it may still be of interest to consider the residuals. All of the residuals are less than .05 and the root mean square of the residuals is .023. Thus, the model provides a reasonably good fit to the data although the model can be confidently rejected statistically.

One possible difficulty with the model in this particular instance is the assumption that the variance of the errors of measurement are constant across time. Judging from the correlation among adjacent grades and the general tendency for measures to be less reliable at the early grades than at the higher grades, one might suspect that $\sigma^2(e_t)$ should be less at grades 4, 5 and 6 than at grades 2 and 3. This problem may contribute to the relatively large residuals in the diagonal at grades 4 and 5.

The second set of data is based on intelligence test scores obtained at 10 points in time for boys at ages 8 through 17. The interval between testing was one year. The correlations which were obtained from Humphreys (1967) were based on data originally collected as part of the Harvard Growth Study. The scores that were intercorrelated are mental age scores. These correlations are reported in section "a" of Table 6-5. Residuals of observed correlations minus correlations estimated from the model are reported in Table 6-5 section "b", and the parameter estimates and chi-square test are reported in Table 6-5 section "c".

The chi-square is again significant. An inspection of the matrix of residuals, however, reveals that the fit is reasonably good with several notable exceptions. The root mean square of the residuals is .035, the largest encountered so far. The magnitude of the root mean square is substantially influenced by a few large residuals. The four largest residuals all involve correlations with scores obtained at age 8. Removing the scores obtained at age 8 would greatly improve the fit. For example, if at age 8 scores were deleted and the remaining variables had the same values of $\hat{R}$, the root mean square residual would be reduced to .026.

PHYSICAL MEASURES

Data were also available for the weight and height of 275 girls obtained on a yearly basis at ages 7 through 16 (Humphreys, 1967). Using the results obtained every second year starting at age 7 an attempt was made to fit these two sets of data to the simplex model.

Table 6-5

Mental Ages of Boys at Various Chronological Ages

a. Intercorrelations

Age	8	9	10	11	12	13	14	15	16	17
8	1.000									
9	.721	1.000								
10	.712	.751	1.000							
11	.747	.721	.816	1.000						
12	.729	.714	.769	.859	1.000					
13	.657	.696	.704	.787	.854	1.000				
14	.598	.634	.726	.745	.778	.864	1.000			
15	.648	.615	.738	.810	.786	.785	.839	1.000		
16	.652	.609	.699	.802	.806	.770	.778	.868	1.000	
17	.556	.588	.604	.736	.775	.780	.750	.778	.848	1.000

b. Residuals ($R-\hat{R}$)

Age	8	9	10	11	12	13	14	15	16	17
8	.000									
9	-.029	.021								
10	.023	-.024	.012							
11	.091	-.017	.004	-.009						
12	.093	-.001	-.018	.013	-.012					
13	.053	.016	-.044	-.018	.021	-.003				
14	.010	-.028	-.003	-.039	-.033	.020	.010			
15	.079	-.025	.033	.052	.002	-.031	.013	-.004		
16	.097	-.015	.012	.064	.042	-.025	-.027	.021	-.015	
17	.028	-.006	-.050	.033	.047	.023	-.017	-.028	.011	.000

Table 6-5 (Continued)

c. Parameter Estimates

Age	Beta	Var(u)
8	—	.864
9	.867	.193
10	.919	.141
11	.952	.100
12	.970	.056
13	.950	.075
14	.974	.033
15	.966	.070
16	.976	.054
17	.952	.068

Var(e) = .135

Chi-Square = 200.98 with 35 df ($p < .001$)

The results are reported in Tables 6-6 and 6-7 for weight and height respectively. In section "a" of each table the intercorrelations are reported. The residuals are reported in section "b" and the parameter estimates and chi-square test are reported in section "c" of each Table.

For both weight and height the chi-square test leads to a rejection of the model.. The residual matrices, however, show a relatively good fit for ages 7, 9, 11, and 13 with a relatively much poorer fit to the correlations involving height or weight at age 15. The estimated variance of the errors of measurement is zero for both height and weight which reflects the high reliability of these physical mesaures but is necessarily an underestimate.

The apparently systematic nature of the residuals for the two sets of physical measures suggests that the simplex model is not adequate for these data. In both cases, the fit is exceptionally good for pairs of measures that are close in time but it becomes less and less adequate for pairs of measures that are further separated in time. For weight (Table 6-6) the average residuals for correlations are .098, .030, .013, and point .000 for measurements separated by 3, 2, 1, and 0 intervening measures respectively. A similar, though less pronounced trend can be seen for height (Table 6-7). This pattern of residuals stands in contrast to those that were observed above for the aptitude and achievement data. For example the averages of the absolute values of the residuals for the vocabulary data (Table 6-4) were .021, .015, 023, and .014 for measures with 3, 2, 1, and 0 intervening measures respectively.

DISCUSSION

The above examples illustrate several points: (1) the simplex model appears to provide a reasonably good fit to at least some sets of academic aptitude and achievement data, (2) where the data do not fit the model very well elements of residual matrix may identify particular problem areas, (3) for the physical measures the pattern of the residuals suggests a general inadequacy of the one step model of the simplex. When the fit is judged to be adequate, the simplex model provides a powerful tool for estimating characteristics of the unobserved error free measures as well as growth statistics of interest.

Table 6-6

Weight of 275 Girls at Various Chronological Ages

a. Intercorrelations

Age	7	9	11	13	15
7	1.000				
9	.880	1.000			
11	.810	.906	1.000		
13	.755	.840	.921	1.000	
15	.744	.773	.790	.880	1.000

b. Residuals ($R-\hat{R}$)

Age	7	9	11	13	15
7	.000				
9	.000	.000			
11	.013	.000	.000		
13	.021	.006	.000	.000	
15	.098	.039	.020	.000	.000

c. Parameter Estimates

Age	Beta	Var(u)
7	--	1.000
9	.880	.225
11	.906	.179
13	.921	.151
15	.880	.225

Var(e) = .000

Chi-Square = 40.92 with 5 df ($p < .001$)

Table 6-7

Standing Height of 275 Girls at Various Chronological Ages

a. Intercorrelations

Age	7	9	11	13	15
7	1.000				
9	.980	1.000			
11	.920	.954	1.000		
13	.887	.909	.923	1.000	
15	.836	.844	.790	.901	1.000

b. Residuals ($R-\hat{R}$)

Age	7	9	11	13	15
7	.000				
9	.000	.000			
11	-.015	.000	.000		
13	.024	.028	.000	.000	
15	.058	.051	-.042	.000	.000

c. Parameter Estimates

Age	Beta	Var(u)
7	—	1.000
9	.980	.039
11	.954	.090
13	.923	.148
15	.902	.188

Var(e) = .000

Chi-Square = 122.34 with 5 df ($p < .001$)

REFERENCES

Anderson, J. E. The limitations of infant and preschool tests in the measurement of intelligence. Journal of Psychology, 1939, 8, 351-379.

Bracht, G. H. & Hopkins, K. D. Stability of educational achievement. In Bracht, G. H., Hopkins, K. D. & Stanley, J. C. (eds.) Perspectives in Educational Measurement, Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Corballis, M. C. Practice and the simplex. Psychological Review, 1965, 22, 399-406.

Cronbach, L. J. & Furby, L. How we should measure "change" -- or should we? Psychological Bulletin, 1970, 74, 68-80.

Guttman, L. A new approach to factor analysis: the radex. In Lazarsfeld, P. L. (ed.) Mathematical Thinking in the Social Sciences, Glencoe, Illinois: Free Press, 1954.

Humphreys, L. G. Investigations of the simplex. Psychometrika, 1960, 25, 313-323.

Humphreys, L. G. The fleeting nature of college academic success. Journal of Educational Psychology, 1968, 59, 375-380.

Humphreys, L. G. Problems in personnel research. In A. L. Fortuna (ed.) Personnel Research and Systems Development, The Personnel Research Laboratory, United States Air Force, Lackland Air Force Base, Texas, 1967, pp. 67-75.

Humphreys, L. G. The misleading aptitude-achievement distinction. Proceedings for the February, 1973 Invitational Conference on the Aptitude-Achievement Distinction, Monterey, California, CTB/McGraw Hill, 1973.

Jones, M. B. Practice as a process of simplification, Psychological Review, 1962, 27, 145-162.

Jöreskog, K. G. Estimation and testing of simplex models. The British Journal of Mathematical and Statistical Psychology, 1970, 23, 121-145.

Jöreskog, K. G., Gravaeus, & van Thillo, M. ACOVS: A general computer program for the analysis of covariance structures. RB 70-15, Princeton, N. J.: Educational Testing Service, 1970.

Jöreskog, K. G. & van Thillo, M. LISREL: A general computer program for estimating a linear structural equation system involving multiple indicators of unmeasured variables. RB-72-56, Princeton, N. J.: Educational Testing Service, 1972.

Linn, R. L. & Werts, C. E. Measurement error in regression. In Walberg, H. J. (ed.) Behavioral Data Analysis, in preparation.

Tatsuoka, K. K. Vector-geometric and Hilbert-space reformulations of classical test theory. Doctoral dissertation, University of Illinois, 1975.

Werts, C. E., Jöreskog, K. G. & Linn, R. L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972, 32, 655-678.

Werts, C. E., Linn, R. L. & Jöreskog, K. G. A simplex model for analyzing academic growth, Educational and Psychological Measurement, in press, a.

Werts, C. E., Linn, R. L. & Jöreskog, K. G. Reliability of college grades from longitudinal data, Educational and Psychological Measurement, in press, b.

CHAPTER 7

CONSTANCY OF CONSTRUCT VALIDITY OVER TIME

Whenever test scores are compared over time the extent to which they are measures of a single common dimension is of concern. This is obviously true when the level of the test is changed and is a prerequisite for vertical equating. Hence, the concerns of this section are closely tied to those that are discussed in the chapter of this report on vertical equating. Even where the same form of a test is used at all times, however, it is possible that different traits are measured by the test at different points in time. An example of such a test might be one that measures problem solving skill at one age and memory or computational accuracy at a later age.

The problem of deciding what is measured by an instrument is basically a problem of construct validity. An important issue for longitudinal studies is the extent to which measures get at the same underlying constructs in a constant fashion over time. If this formulation is accurate, then all of the procedures and considerations involved in the ongoing task of construct validation would apply to the concerns of longitudinal measures of change. Thus, the variety of correlational, experimental, and logical procedures discussed by Cronbach (1971) are relevant when attempts are made to measure the same trait at two or more points in time. But, the problem is complicated by the addition of the time dimension.

PATTERN OF INTERCORRELATIONS

When plotting trends or calculating change scores it is typically assumed that the same thing is being measured at each point in time. From the observation that scores change from one test administration to the next, however, it is not clear whether the people have changed along a given dimension or what is measured by the test has changed.

> "If the correlation between pretest and posttest is reasonably high, we are inclined to ascribe change scores to changes in the individuals. But if the correlation is low, or if the pattern of correlations with other variables is different on the two occasions, we may suspect that the test does not measure the same thing on the two occasions. Once it is allowed that the pretest and posttest measure different things, it becomes embarrassing to talk about change (Bereiter, 1963, p. 11)."

Bereiter's comments suggest that the pattern of correlations of the focus variable with other variables is highly relevant as evidence that the measures are getting at the same thing. Although this contention is closely related to the approach that is discussed below, it must be acknowledged at the outset that even the existence of identical

correlations of the focus variable with a host of other variables would not guarantee that the same thing is being measured. At best, the similarity of the pattern of correlations can improve the plausibility of the claim that the same thing is being measured by making alternative explanations seem less likely. The logical difficulty of concluding that similar correlations imply measurement of the same dimension is easily ignored.

Suppose, for example, that at time 1 measure X_1 correlates .35, .15 and .18 with measures X_2, X_3 and X_4 respectively. At time 2 the correlations of X_1 with X_2, X_3 and X_4 are .51, .44 and .49 respectively. These results might lead to a suspicion that measure X_1 was measuring somewhat different things but that is not necessarily the case. In fact, both sets of correlations were derived from the same model with two latent traits. At both points in time it was assumed that each X_{jt} was a linear function of two latent traits, Z_1 and Z_2, and an uncorrelated error of measurement, e_{jt}, where j indexes the measures and t indexes the time of measurement.

More formally the model that was used to derive the correlations at a particular point in time can be expressed

$$\underset{\sim}{X} = \underset{\sim}{\mu} + B \underset{\sim}{Z} + \underset{\sim}{e} \qquad (1)$$

where $\underset{\sim}{X}$ is a column vector of observations on the p observed variables, $\underset{\sim}{\mu}$ is a column vector of p means, $\underset{\sim}{Z}$ is a column vector of scores on the k latent traits, B is a p by k matrix of weights, and $\underset{\sim}{e}$ is a column vector of errors of measurement on the p measures. It is assumed that the elements in $\underset{\sim}{e}$ are mutually uncorrelated and uncorrelated with the latent traits. The above model is, of course, simply a factor model except that the errors of measurement would normally be replaced by specific factors.

With the above model the variance-covariance matrix among the observed variables is

$$\Sigma = B \Gamma B' + \theta^2 \qquad (2)$$

where Γ is the variance-covariance matrix among the latent traits and θ^2 is a diagonal p x p matrix with the error variances in the diagonal.

Returning to the example of correlations of X_1 with X_2, X_3 and X_4 at time 1 and time 2, the correlations at both points in time were generated with the same B and θ^2 matrices. In both cases B was

$$B = \begin{bmatrix} .7 & 0 \\ .6 & .4 \\ 0 & .6 \\ 0 & .8 \end{bmatrix}$$

and all the error variances were assumed to equal 1.0. At both points in time the variance of Z_1 was also assumed to equal 1.0. Thus, at both points in time

$$X_{1t} = .7\, Z_1 + e_{1t} \quad ,$$

where t refers to time. That is, precisely the same thing is being measured with the same degree of accuracy. Only the variance of Z_2 and the covariance of Z_1 and Z_2 were changed from time 1 to time 2. All observed measures remained the same linear function of two latent traits plus an uncorrelated error of measurement with the same variance and

$$Z_{12} = Z_{11} = Z_1$$

while

$$Z_{22} = 2\, Z_{11} \quad .$$

Without belaboring this admittedly artificial example further the main point is simply the one stated originally. Namely, the similarity of the pattern of correlations of a measure with a variety of other measures at two points in time does not imply whether the same or different things are being measured.

A similar approach to making inferences about the constancy of what is being measured by a variable is to compare standardized factor loadings. If two sets of standardized factor loadings are equal or proportional it is sometimes inferred that the variables are measuring the same things at different points in time. Given the above arguments about intercorrelations, it is hardly surprising that such an inference or its converse based on non-proportional standardized loadings is not justified (see Werts, Jöreskog and Linn, 1972, pp. 673-675).

A better approach to the problem is to compare unstandardized factor weights. If the same latent trait is being measured then the unstandardized factor weights should be constant assuming a linear factor model. This of

course is a strong assumption which may not be justified. Within the model, however, different weight matrices would imply that different things are being measured. Unfortunately, the same B and Γ matrices do not necessarily imply the same factors. Speaking in a slightly different context, McGaw and Jöreskog note that "...there is no mathematical basis for the inference of identity of common factors across populations, even in the case where common... [B and Γ] can be fitted to all populations. It is clearly possible...that identical dispersion matrices could be obtained from different test batteries...(1971, p. 165)." The same statement would apply within our context of the same population measured at two or more points in time.

Although common B and Γ don't conclusively imply the identity of common factors at different points in time it is still of value to be able to reject the proposition that the common factors are the same when the matrices are different. Furthermore, "...the inference of identical factors seems reasonable if the ...[B and Γ] matrices are the same... (McGaw and Jöreskog, 1971, p. 165)". Even if only the B matrices are the same as in the example used above, the same substantive interpretation seems reasonable albeit with different variances and interrelationships among the latent variables.

CONGENERIC MEASURES OVER TIME

A relatively simple yet conceptually appealing model for measures of the same trait over time is provided by the notion of congeneric measures (Jöreskog, 1968, 1971). Except for errors of measurement, congeneric tests measure the same trait and their true scores are linearly related. As applied to the longitudinal situation an observation on measure j at time t, X_{jt}, would be given by

$$X_{jt} = \mu_{jt} + b_{jt} Z_j + e_{jt}$$

where μ_{jt} is the mean, b_{jt} is the weight for variable j at time t, Z_j is the latent variable for variable j, and e_{jt} is the error of measurement on variable j at time t. The lack of a t subscript on the Z corresponds to the assumption that measure j measures the same trait at all points in time. As usual, the errors of measurement are assumed to be mutually uncorrelated and uncorrelated with the latent traits.

Even with only observations on a single measure the hypothesis that the measures are congeneric may be tested assuming multivariate normality providing observations on four or more occasions are available (Jöreskog, 1968). There would still be advantages to having several sets of measures, however, since this would provide a more powerful test of the model, especially the assumption that the error terms in the model are uncorrelated with all other variables. Although the above approach is attractive with measures available at numerous points in time, by far the most typical situation encountered in longitudinal studies is where the same measures are obtained at only two points in time. Also, for most data sets involving measures of academic achievement, the simplex

model discussed in another chapter of this report is apt to provide a better fit. With only two points in time and with only a single measure at each occasion, no test of the model is possible.

With three or more measures available at two points in time, models can be constructed to test whether each measure is congeneric over the two time points. The test would not be specific to this hypothesis alone, however. The model would also involve specifications of the factor structure of the latent trait dispersion matrix, Γ. Following Joreskog (1968, 1971) the factor model for Γ may be specified

$$\Gamma = \Lambda \Phi \Lambda' + \Psi$$

where Λ is a matrix of factor loadings for true scores, Φ is the variance-covariance matrix among the factors underlying the true scores and Ψ is a diagonal matrix of uniquenesses. With this structure of Γ the full model may be expressed

$$\Sigma = B(\Lambda \Phi \Lambda' + \Psi) B' + \theta^2$$

which may be analyzed following procedures described in Joreskog (1970).

To illustrate this approach two small examples each involving three tests with scores at two points in time were selected.

Example 1: For the first example data on two arithemetic tests and an attitudinal measure were used. These measures were used for 75 children before and after an instructional program in arithmetic. The variance-covariance matrix for the 6 variables is reported in Table 1.

The model specified that a given measure at two points in time is congeneric and that there is one common and three specific factors underlying the three true scores. Thus, with the tests ordered tests 1, 2, and 3 at time 1 then tests 1, 2, and 3 at time 2 as they are for the variance-covariance matrix in Table 1, the model specifies that the B matrix will have four zeros and two values to be estimated in each column. The pattern is

$$B = \begin{bmatrix} * & 0 & 0 \\ 0 & * & 0 \\ 0 & 0 & * \\ * & 0 & 0 \\ 0 & * & 0 \\ 0 & 0 & * \end{bmatrix}$$

where the asterisks are the values to be estimated.

TABLE 7-1

Variance-Covariance Matrix

(Example 1, N = 75)

Variable		1	2	3	1	2	3
	Time	1	1	1	2	2	2
1. Arith. 1	1	118.50					
2. Arith. 2	1	45.33	46.68				
3. Attitude	1	257.46	135.38	2555.80			
1. Arith. 1	2	73.66	39.82	239.62	94.00		
2. Arith. 2	2	56.99	39.40	149.62	48.29	58.17	
3. Attitude	2	238.21	126.62	1166.40	159.15	152.25	1683.00

The Λ matrix has three rows and one column with all entries free, Φ is just a scaler of 1.0, Ψ is a 3 x 3 diagonal matrix with all diagonal entries free. The maximum likelihood solution for the variance-covariance matrix of example 1 is presented in Table 2.

All three variables have substantial weights on the general factor. For each variable the weights in $\hat{B}$ are reasonably similar at the two points in time. The attitude measure has an apparently large variance of the errors of measurement but the true score variance is also very large on this variable in comparison to the other two variables. The critical question regarding the above results is the adequacy of the model for the data. This is answered in two ways: by a chi-square test of fit and by an inspection of the matrix of residuals. The chi-square for these data is 5.95 with 3 degrees of freedom which is not significant at the .10 level.

The matrix of residuals, i.e., the observed variance-covariance matrix minus the variance-covariance matrix estimated by the model is reported in Table 3. The residuals shown in Table 3 are generally small compared to the corresponding elements in Table 1. The largest residual not only in absolute magnitude but as a ratio of the corresponding element in Table 1 is for the covariance of variable 1, time 2 with variable 3, time 2. All of the larger residuals involve variable 3 which may not be surprising given that variable 3 is an attitude measure whereas the other two are achievement tests.

Although the above model provides a reasonably satisfactory fit it is not a very severe test of the hypothesis that each measure measures the same thing at both points in time. A total of 18 parameters (6 in B, 6 in θ^2 , 3 in Λ, and 3 in Ψ) were estimated from a total of only 21 distinct elements in the observed variance-covariance matrix. A more severe test would be provided with more measures, more points in time or fewer parameters. One way to reduce the number of parameters is to make the model more restrictive. For example, the variance of the errors of measurement of a given measure might be assumed to be equal at both points in time. This would reduce the number of parameters to be estimated in θ^2 from 6 to 3 and require a total of 15 rather than 18 parameters to be estimated.

With the equal error variance restraint added, the parameter estimates reported in Table 4 were obtained for the variance-covariance matrix

TABLE 7-2

Maximum Likelihood Solution (Example 1)

		$\hat{B}$ Matrix			Diagonal Entries
j	t	1	2	3	in $\hat{\theta}^2$
1	1	3.30	.0*	.0	5.50
2	1	.0	1.51	.0	3.82
3	1	.0	.0	6.32	34.71
1	2	2.76	.0	.0	5.69
2	2	.0	1.86	.0	3.12
3	2	.0	.0	5.46	26.00

j	$\hat{\Lambda}$ Matrix	Entries in $\hat{\Psi}$
1	2.85	.00
2	3.27	1.82
3	4.38	3.82

*Fixed by hypothesis

TABLE 7-3

Residual Matrix (Example 1)

j		1	2	3	1	2	3
	t	1	1	1	2	2	2
1	1	.00					
2	1	-1.16	.00				
3	1	-2.70	-1.69	.01			
1	2	- .07	- .99	22.29	.00		
2	2	- .15	.00	-18.79	.56	.00	
3	2	13.65	8.34	- .01	-28.45	6.88	.00

TABLE 7-4

Maximum Likelihood Solution With Constant

Error Variances for Each Measure (Example 1)

		$\hat{B}$ Matrix			Entries in $\hat{\theta}^2$
j	t	1	2	3	
1	1	2.54	.0*	.0	5.59**
2	1	.0	1.97	.0	3.55
3	1	.0	.0	5.68	29.59
1	2	2.16	.0	.0	5.59
2	2	.0	2.31	.0	3.55
3	2	.0	.0	4.15	29.59

j	$\hat{\Lambda}$ Matrix	Entries in $\hat{\Psi}$
1	3.68	.00
2	2.57	1.42
3	5.20	4.82

* Fixed by hypothesis

** The pairs of elements 1 and 4, 2 and 5, and 3 and 6 are restrained to be equal.

in Table 1. The solution shown in Table 4 yields a chi-square of 8.07 with 6 degrees of freedom which is not significant at the .20 level. While the resulting residuals are slightly larger than those shown in Table 3, the model even with the restriction of equal error variance for a given measure at the two points in time appears reasonable.

A still more restrictive model for the above data is provided by requiring that not only the error variances but the entries in B be the same for a given measure at the two points in time. This is equivalent to the hypothesis that each measure at time 2 is parallel to the corresponding measure at time 1 except for a possible additive constant from time 1 to time 2. This is a very restrictive model for longitudinal measures. It says, in effect, that the only two possible differences between time 2 and time 1 measures are different means and different errors of measurement. The underlying true scores are identical within an additive constant and the errors of measurement are uncorrelated and have equal variances. With these additional restrictions the estimates reported in Table 5 were obtained.

The chi-square for the rather highly restricted solution shown in Table 5 is 16.33 which with 9 degrees of freedom (21 separate elements in the variance-covariance matrix minus 12 parameters to be estimated) has an associated p value of .06. Although not significant, this increase in the chi-square suggests that the model may be too restrictive. A test of the additional restriction of equal regression weights is provided by the difference in the chi-squares associated with the solutions in Tables 4 and 5. This difference is 8.26 and with 3 degrees of freedom is significant at the .05 level. This suggests that the restriction of equal entries in B is not reasonable.

Example 2: As a second example, data available on three arithmetic subtests (subtraction, multiplication, and division) at two points in time were used. The variance-covariance matrix for a sample of 47 fourth grade students on these six variables is shown in Table 6. The maximum likelihood solution for the model specifying congeneric measures over time and one factor underlying the true scores is shown in Table 7. The chi-square test of the model is 11.52 which with 3 degrees of freedom is significant at the .01 level. Thus, in contrast to the results for example 1, the least restrictive model can be confidently rejected for the data in example 2.

Part of the problem with the model may be suggested by the entries in the residual matrix which is shown in Table 8. Three of the four largest residuals all involve the multiplication test. It may be that the hypothesis that a test is congeneric over time is least reasonable for the multiplication test.

LESS RESTRICTIVE MODELS

As was previously indicated, the hypothesis of congeneric measures over time may be much too restrictive in most longitudinal situations. The notion of growth does not normally involve the strong assuption that the true score at time t is merely a linear function of the true score at

TABLE 7-5

Maximum Likelihood Solution with Constant Error Variance and Regression of Observed on True Score for Each Measure (Example 1)

		$\hat{B}$ Matrix			Entries in $\hat{\theta}^2$
j	t	1	2	3	
1	1	2.67*	.0**	.0	5.69*
2	1	.0	2.47	.0	3.61
3	1	.0	.0	5.79	30.87
1	2	2.67	.0	.0	5.69
2	2	.0	2.47	.0	3.61
3	2	.0	.0	5.79	30.87

j	$\hat{\Lambda}$ Matrix	Entries in $\hat{\Psi}$
1	3.22	.00
2	2.24	1.20
3	4.46	3.85

* Pairs of entries for a given measure are restrained to be equal

** Fixed by hypothesis

TABLE 7-6

Variance-Covariance Matrix

(Example 2, N = 47)

	Variable		1	2	3	1	2	3
		Time	1	1	1	2	2	2
1	Subtraction	1	2.35					
2	Multiplication	1	1.24	2.54				
3	Division	1	.63	.58	2.47			
1	Subtraction	2	.70	.39	.10	1.56		
2	Multiplication	2	.96	.41	.71	.93	2.52	
3	Division	2	1.52	.95	1.02	1.10	1.83	3.37

TABLE 7-7

Maximum Likelihood Solution (Example 2)

j	t	$\hat{B}$ Matrix 1	2	3	Entries in $\hat{\theta}^2$
1	1	.92	.0*	.0	1.17
2	1	.0	.59	.0	1.46
3	1	.0	.0	.46	1.45
1	2	.67	.0	.0	1.02
2	2	.0	1.01	.0	1.12
3	2	.0	.0	1.31	.70

j	$\hat{\Lambda}$ Matrix	Entries in $\hat{\Psi}$
1	1.07	.00
2	1.11	.00
3	1.20	.50

* Fixed by hypothesis

TABLE 7-8

Residual Matrix (Example 2)

j		1	2	3	1	2	3
	t	1	1	1	2	2	2
1	1	.00					
2	1	.60	.00				
3	1	.08	.22	.00			
1	2	-.01	-.08	-.30	.00		
2	2	-.15	-.32	.09	.12	.00	
3	2	-.03	-.07	.00	-.02	.06	.00

time t - 1. Rather, we would normally like to assume that the rank order of individuals along a given dimension may change over time. Once the rank order on the underlying dimension is allowed to change, however, there is a difficulty in establishing whether it is the trait being measured or the people that are changing. Thus, the fundamental problem with which we started this chapter still remains.

If a complete model can be specified it may sometimes be tested within the context of the general procedures for the analysis of covariance structures (Jöreskog, 1970). In most instances, the theory is apt to be lacking to make this more than an approach to testing the reasonableness of a variety of possibilities. With three or more occasions and several measures the procedures described by Jöreskog (1969) for factoring a multitest-multioccasion matrix should be of value. When restricted to two points in time as is typically the case, however, strong assumptions about the causal structure of the unmeasured variables are apt to be needed.

An approach to the problem involving multiple measures of a trait at time 1 and again at time 2 as well as multiple measures of a second variable that is thought to be a determinant of growth is discussed by Werts, et al. (1972). While potentially useful, their approach makes heavy practical demands for a closed model with all intercorrelated determinates for final status on the trait of interest included. It also requires multiple measures (at least three) of each trait.

Several attempts were made to illustrate the approach described in Werts, et al. using Project TALENT results reported by Shaycoft (1967). We were not successful, however, in finding examples for which the fit was good enough to provide useful illustrations of the approach. This failure is probably due, in large part, to the artificial nature of the examples that were attempted. The Project TALENT data collection was not designed with such an analytical model in mind and the needed multimethod approach to the measurement of each trait was not used. As a result the identification of "methods" factors and of a causal model for analysis were too crude to be successful.

CONCLUSIONS

The problem of deciding if it is the people or the nature of the dimension that is changing is basically a problem of construct validity. As such, it is an unending process for which theory, logical analysis and a variety of empirical procedures are relevant. Assuming linearity, the procedures for the analysis of covariance structures (Jöreskog, 1970) provide a potentially powerful analytical tool in this effort. But, there are two major obstacles to the application of this approach. These are the lack of theory to guide the testing of specific hypotheses and the requirement of multiple measures for all but the simplest of hypotheses.

REFERENCES

Bereiter, C. Some persisting dilemmas in the measurement of change. (in C. W. Harris, ed.), Problems in Measuring Change. Madison, Wisconsin, University of Wisconsin Press, 1963, 3-20.

Cronbach, L. J. Test validation. (In R. L. Thorndike, ed.) Educational Measurement, second edition. Washington, D. C.: American Council on Education, 1971, 443-507.

Jöreskog, K. G. Statistical models for congeneric test scores. Proceedings of the American Psychological Association, 76th Annual Convention, 1968, 213-214.

Jöreskog, K. G. Factoring the multitest-multioccasion matrix. Princeton, New Jersey: Educational Testing Service, Research Bulletin, 69-62, 1969.

Jöreskog, K. G. A general method for the analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Jöreskog, K. G. Statistical analysis of sets of congeneric tests. Psychometrika, 1971, 36, 109-133.

Jöreskog, K. G., Gruvaeus, G. T., & van Thillo, M. ACOVS: A general computer program for the analysis of covariance structures. Princeton, New Jersey: Educational Testing Service, Research Bulletin 70-15, 1970.

McGaw, B. & Joreskog, K. G. Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. British Journal of Mathematical and Statistical Psychology, 1971, 24, 154-168.

Shaycoft, M. F. The High School Years: Growth in Cognitive Skills. Unpublished Technical Report. Pittsburgh: The American Institutes for Research, 1967.

Werts, C. E., Jöreskog, K. G. & Linn, R. L. A multitrait-multimethod model for studying growth. Educational and Psychological Measurement, 1972, 32, 655-678.

Chapter 8

TIME-SERIES ANALYSIS APPLIED TO LONGITUDINAL STUDIES

INTRODUCTION

Time-series analysis refers to the body of knowledge and techniques that deals with the fitting of stochastic models to a series of observations made at successive, equally spaced time points. It thus differs from techniques for fitting deterministic models such as polynominal and multiple regression equations. Developed primarily in the context of industrial engineering, economics, and business management, its primary purpose heretofore has been forecast and control. (Box and Tiao, 1965; Box and Jenkins, 1970; Nelson, 1973.)

The application of time-series analysis to behavioral and social sciences in general, and to educational and psychological research in particular, has been pioneered by Campbell (1969) and Glass, Willson and Gottman (1975), among others. The main objective of these works has been the application of the technique to "interrupted time-series experiments," i.e., studies in which series of observations both before and after the introduction of some experimental intervention are involved, and whose aim is to examine the nature and significance of the effects of the intervention, if any.

The purpose of this chapter are threefold. First, to present a more elementary exposition of the methodology of time-series analysis than is available in the literature to date; second, to point out that, as currently used, the method does not take into account the longitudinal nature of the data, but rather treats them as sequential

cross-sectional data; third, to suggest some modifications to make the technique specifically applicable to genuine longitudinal studies.

THE BASIC MODELS

Within the rubric of linear models, the most general stochastic model for discrete time-series observations is one which postulates that the observation z_t at time t is expressible as a linear combination of an overall "level" parameter L and random disturbances (or white noise) at time t and all prior time points, a_t, a_{t-1}, a_{t-2} ... . That is,

$$[1] \qquad z_t = L + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots$$

which is called the general discrete linear stochastic process model, or the "linear filter" model for short. In order to achieve anything resembling tractability, we must assume that the random disturbances a_t are identically and independently distributed random variables with mean 0 and variance σ_a^2. For inferential purpose we further assume that the common distribution is normal; i.e.,

$$a_t \sim \text{IND}\ (0, \sigma_a^2)$$

At first glance it may seem that for any stochastic process expressible by Eq. [1], it should follow that,

$$\begin{aligned} E(z_t) &= L + E(a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots) \\ &= L + E(a_t) + \psi_1 E(a_{t-1}) + \psi_2 E(a_{t-2}) + \cdots \\ &= L + 0 + 0 + 0 + \cdots \\ &= L. \end{aligned}$$

This fallacious, however, in that the transition from the first to the second step is not valid unless the infinite series $a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots$ is convergent. The necessary and sufficient condition for this to be the case is that the coefficient series, $\sum_{i=0}^{\infty} \psi_i$ (where $\psi_o = 1$), itself be convergent. If, and only if, this is true, we can assert that $E(z_t) = L$ for all t. Thus, as a first principle, we have:

[2] $E(z_t) = L$, for all t, iff $\sum_{i=0}^{\infty} \psi_i = K < \infty$.

When this condition holds, process [1] is said to be <u>stationary through the second moments</u>, for as we shall immediately see, the condition also implies that the variance $Var(z_t)$ and covariances between staggered z_t's are independent of t. Together with the normality assumption for the distribution of a_t, stationarity through the second moments assures <u>complete stationarity</u>--i.e., that the probability distribution of z_t is invariant with respect to t. Intuitively, a stationary process is one in which the successive observations, although "meandering" in time, always centers around a fixed mean, $E(z_t) = L$.

Let us now verify the above assertion that the condition stipulated in [2] is sufficient also to guarantee that $Var(z_t)$ exists and is independent of t.

$$Var(z_t) = E(z_t - L)^2$$

$$= E(a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots)^2$$

$$= E(a_t^2 + \psi_1^2 a_{t-1}^2 + \psi_2^2 a_{t-2}^2 + \ldots)$$

$$+ 2E(\psi_1 a_t a_{t-1} + \psi_2 a_t a_{t-2} + \ldots$$

$$+ \psi_1 \psi_2 a_{t-1} a_{t-2} + \ldots)$$

$$= \sigma_a^2(1+\psi_1^2+\psi_2^2+\ldots),$$

since $E(a_t a_{t'}) = 0$ for $t \neq t'$ because the a_t are assumed to be independently distributed. Obviously, the convergence of $\sum_{i=0}^{\infty} \psi_i$ assures $\sum_{i=0}^{\infty} \psi_i^2$ also to be convergent. We have thus shown that

[3] $$\text{Var}(z_t) = \sigma_a^2 \sum_{i=0}^{\infty} \psi_i^2, \text{ for all } t, \text{ iff } \sum_{i=0}^{\infty} \psi_i = K < \infty.$$

Similarly, it can be shown that

[4] $$\text{Cov}(z_t, z_{t+j}) = \alpha_a^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+j}, \text{ iff } \sum_{i=0}^{\infty} \psi_i = K < \infty.$$

In the literature of time-series analysis, $\text{Var}(z_t)$ for stationary processes is denoted by γ_o and $\text{Cov}(z_t, z_{t-j})$ by γ_j, the latter being called the autocovariance of lag j.

A simple example of a stationary process is one for which the coefficients ψ_i in [1] are given by

$$\psi_i = \phi^i, \text{ where } |\phi| < 1.$$

In this case,

$$\sum_{i=0}^{\infty} \psi_i^2 = 1 + \phi^2 + \phi^4 + \phi^6 + \ldots = \frac{1}{1-\phi^2}$$

and

$$\sum_{i=0}^{\infty} \psi_i \psi_{i+j} = \phi^j + \phi^{j+2} + \phi^{j+4} + \ldots = \frac{\phi^j}{1-\phi^2}.$$

Hence, Eqs. [3] and [4] specialize to

[3*] $$\gamma_o = \sigma_a^2/(1-\phi^2)$$

and

[4*] $\gamma_j = \sigma_a^2 \phi^j / (1-\phi^2)$.

Moving-Average Processes

An even simpler way in which Eq. [1] can represent a stationary process is when the coefficients ψ_i are all zero for $i > q$. The series $a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots$ then terminates with the term $\psi_q a_{t-q}$, and the coefficient series $\sum_{i=0}^{\infty} \psi_i = \sum_{i=0}^{q} \psi_i$ necessarily converges. The resulting process,

[5] $z_t = L + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots + \psi_q a_{t-q}$

is called a moving-average process of order q, abbreviated MA(q).[1]

---

[1]The phrase "moving-sverage" does not mean that the average of z_t "moves" or varies with t--otherwise the process would be non-stationary. It simply means that $z_t - L$ is a weighted composite of the set of disturbances through q time points back, which of course moves with t. For example, with $q = 2$, $z_5 - L$ is a weighted composite of a_5, a_4 and a_3; $z_{10} - L$ is a weighted composite of a_{10}, a_9 and a_8. It is the set of a's of which z_t is a weighted composite that moves with t. Note also that the weighted composite, $a_t + \psi_1 a_{t-1} + \ldots + \psi_q a_{t-q}$, is not really a weighted average, since the coefficient 1, ψ_1, ψ_2, ..., ψ_q do not, in general, sum to unity [as Box and Jenkins (1970, p. 10) points out]. For historical reasons, the phrase "moving-average" is retained even though it is, strictly speaking, a misnomer.

For purely historical reasons again, the coefficients ϕ_i ($i>0$) are replaced by $-\theta_i$, so the conventional equation for an MA(q) process is

[6] $$z_t = L + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q}$$

Thus, the simplest case (which turns out to be adequate for many situations) is written as

[7] $$z_t = L + a_t - \theta_1 a_{t-1}$$

For MA(1), it follows from Eqs. [3] and [4] that

[8] $$\gamma_0 = \sigma_a^2(1+\theta_1^2)$$

and

[9] $$\gamma_1 = \sigma_a^2(-\theta_1)$$

while

[10] $$\gamma_j = 0 \text{ for } j > 1.$$

In addition to the variance and autocovariances of various lags, another important parameter for stationary time-series models is the autocorrelation of lag j. Its importance lies in the fact that its sample counterpart is one of the main statistics used for identifying the appropriate model for a given series of observed data, as we shall see later. The autocorrelation of lag j, denoted by ρ_j, is computed in the usual way, as

$$\rho_j = \frac{\text{Cov}(z_t, z_{t+j})}{\sqrt{\text{Var}(z_t)}\sqrt{\text{Var}(z_{t+j})}}$$

But, since $Cov(z_t, z_{t+j}) = \gamma_j$ and $Var(z_t) = Var(z_{t+j}) = \gamma_o$, ρ_j may be expressed as

[11] $$\rho_j = \gamma_j/\gamma_o.$$

Thus, for MA(1) we have

[12] $$\rho_1 = -\theta_1/(1+\theta_1^2) \text{ and } \rho_j = 0 \text{ for } j > 1.$$

In general, for MA(q) the autocorrelations of lags less than or equal to q are non-zero, and those of lags greater than q are zero. For instance, for MA(2) we have Eqs. [3] and [4]

$$\gamma_o = \sigma_a^2(1+\theta_1^2+\theta_2^2)$$

$$\gamma_1 = \sigma_a^2(-\theta_1+\theta_1\theta_2)$$

$$\gamma_2 = \sigma_a^2(-\theta_2).$$

Hence,

[13] $$\rho_1 = (-\theta_1+\theta_1\theta_2)/(1+\theta_1^2+\theta_2^2)$$

$$\rho_2 = -\theta_2/(1+\theta_1^2+\theta_2^2)$$

$$\rho_j = 0 \qquad \text{for } j > 2.$$

Autoregressive Processes

Another important class of processes is the autoregressive process (AR). The equation for AR is obtained by going back to the general linear filter of Eq. [1] and rewriting the right-hand side in terms of the current disturbance and all past observations. To do so, we first transpose the terms in Eq. [1] to get

$$a_t = z_t - L - \psi_1 a_{t-1} - \psi_2 a_{t-2} - \cdots$$

and, noting that this holds for any time point, we, have, e.g. for t-1,

$$a_{t-1} = z_{t-1} - L - \psi_1 a_{t-2} - \psi_2 a_{t-3} - \cdots .$$

substituting this in Eq. [1] in its original form, we get

$$z_t = L + a_t + \psi_1(z_{t-1} - L - \psi_1 a_{t-2} - \psi_2 a_{t-3} - \ldots) + \psi_2 a_{t-2} + \cdots$$

$$= L(1-\psi_1) + \psi_1 z_{t-1} + a_t + (\psi_2 - \psi_1^2) a_{t-2} + (\psi_3 - \psi_1\psi_2) a_{t-3} + \cdots$$

from which a_{t-1} has been eliminated. Similarly, we may successively eliminate a_{t-2}, a_{t-3}, etc., and ultimately get an equation of the form

[14] $$z_t = \beta + \pi_1 z_{t-1} + \pi_2 z_{t-2} + \cdots + a_t,$$

where the coefficients π_i are functions of the ψ_i's and the constant β is a function of L and the ψ_i's. The name "autoregressive model" comes from the fact that Eq. [14] resembles a multiple regression equation with z_t as the criterion variable, the past observations z_{t-1}, z_{t-2} ... as predictors, and a_t as the error of estimate.

Of course the series $\pi_1 z_{t-1} + \pi_2 z_{t-2} + \ldots$ must converge before [14] has any chance of representing a stationary process. But, as we shall see below, such convergence is only a necessary but insufficient condition for stationarity. As before, the simplest way to assure convergence of the series is to require that all the coefficients beyond the p^{th}, say, shall vanish. When this is the case, we have an autoregressive process of order p, symbolized AR(p). Again for historical reasons, the coefficients π_i are rewritten as ϕ_i, and the conventional equation for AR(p) is

[15] $$z_t = \beta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + a_t.$$

For the simplest case, AR(1), we have

[16] $$z_t = \beta + \phi_1 z_{t-1} + a_t.$$

It may be tempting to take the expected values of both sides of this equation to get

$$E(z_t) = \beta + \phi_1 E(z_{t-1}) + 0,$$

and, letting $E(z_{t-1}) = E(z_t)$, obtain

$$E(z_t) = \beta/(1-\phi_1).$$

However, this already assumes the process to be stationary [when we put $E(z_{t-1}) = E(z_t)$], whereas in fact it may not be. To see why Eq. [16] does not automatically represent a stationary process despite its having only two variable terms on the right, we must convert the equation back to MA form--i.e., a linear combination of present and past disturbances--for which we already know the condition for stationarity.

This is done by using [16] with t replaced by t-1 to express z_{t-1} in terms of z_{t-2} and a_{t-1}, as

$$z_{t-1} = \beta + \phi_1 z_{t-2} + a_{t-1},$$

whence

$$z_t = \beta + \phi_1(\beta + \phi_1 z_{t-2} + a_{t-1}) + a_t$$

$$= (1+\phi_1)\beta + \phi_1^2 z_{t-2} + a_t + \phi_1 a_{t-1};$$

then z_{t-2} is expressed in terms of z_{t-3} and a_{t-2}, and so forth. Continuing in this vein, we eventually get

$$z_t = \beta(1+\phi_1+\phi_1^2+\ldots) + a_t + \phi_1 a_{t-1} + \phi_1^2 a_{t-2} + \ldots .$$

Thus, both the series in the a_t's and the series forming the multiplier of β converge if and only if $|\phi_1| < 1$. Once this condition is met, this equation is seen to be equivalent to

$$[17] \qquad z_t = \frac{\beta}{1-\phi_1} + \sum_{i=0}^{\infty} \phi_1^i a_{t-i},$$

which is precisely the process we referred to earlier as an example of a moving-average process of infinite order which nevertheless is stationary; $\beta/(1-\phi_1)$ here plays the role of L. Thus, an AR(1) process is, under the condition stated, equivalent to an MA process of infinite order. We thus conclude that, if and only if $|\phi_1| < 1$, Eq. [16] represents a stationary process with

$$[18] \qquad E(z_t) = \beta/(1-\phi_1),$$

and, from Eqs. [3*] and [4*],

$$\gamma_j = \sigma_a^2 \phi_1^j/(1-\phi_1^2) \quad (j = 0,1,2,\ldots).$$

Consequently, the autocorrelation of lag j is

$$[19] \qquad \rho_j = \gamma_j/\gamma_0 = \phi_1^j.$$

Unlike for a MA process, the autocorrelation does not suddenly vanish after a certain lag, but steadily decreases exponentially.

The equation for AR(1) is often written in deviation-score form, thus: let

$$z_t - E(z_t) = a_t - \frac{\beta}{1 - \phi_1} = \tilde{z}_t$$

Then, from Eq. [16],

$$z_t = (\beta + \phi_1 z_{t-1} + a_t) - \frac{\beta}{1 - \phi_1}$$

$$= \beta \frac{-\phi_1}{1 - \phi_1} + \phi_1 z_{t-1} + a_t$$

$$= \phi_1 (z_{t-1} - \frac{\beta}{1 - \phi_1}) + a_t$$

$$= \phi_1 \tilde{z}_{t-1} + a_t$$

Thus, the equation for AR(1) in deviation-score form,

[20] $$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + a_t$$

is the same as [16] except for the absence of the constant term β.

AR Processes of Order Two and Higher. The model equation for AR(2) is

[21] $$z_t = \beta + \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t.$$

Once it is ascertained that the stationarity condition (to be specified later) is satisfied, we may get $E(z_t)$ by taking the expected values of both sides of [21], letting $E(z_{t-1}) = E(z_{t-2}) = E(z_t)$ and solving to obtain

[22] $$E(z_t) = \beta/(1-\phi_1-\phi_2).$$

To compute the variance and autocovariances it is convenient to use the deviation-score form of Eq [21]:

[23] $$z_t = \phi_1 \tilde{z}_{t-1} + \phi_2 \tilde{z}_{t-2} + a_t,$$

where $$\tilde{z}_t = z_t - \beta/(1-\phi_1-\phi_2).$$

Then

$$\gamma_o = E(\tilde{z}_t^2) = E[\tilde{z}_t(\phi_1\tilde{z}_{t-1}+\phi_2\tilde{z}_{t-2}+a_t)]$$

$$= \phi_1 E(\tilde{z}_t\tilde{z}_{t-1}) + \phi_2 E(\tilde{z}_t\tilde{z}_{t-2}) + E(\tilde{z}_t a_t)$$

$$= \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma_a^2$$

The last term in the last step obtains because, from Eq. [23],

$$E(\tilde{z}_t a_t) = E[(\phi_1\tilde{z}_{t-1}+\phi_2\tilde{z}_{t-2}+a_t)a_t]$$

$$= \phi_1 E(\tilde{z}_{t-1}a_t) + \phi_2 E(\tilde{z}_{t-2}a_t) + E(a_t^2),$$

and observations prior to time t one, of course, independent of the disturbance a_t at time t. Similarly,

$$\gamma_1 = E(\tilde{z}_t\tilde{z}_{t-1}) = E[(\phi_1\tilde{z}_{t-1}+\phi_2\tilde{z}_{t-2}+a_t)\tilde{z}_{t-1}]$$

$$= \phi_1\gamma_o + \phi_2\gamma_1,$$

and

$$\gamma_2 = E(\tilde{z}_t\tilde{z}_{t-2}) = \phi_1\gamma_1 + \phi_2\gamma_o.$$

We thus have the set of equations

$$[24] \quad \begin{cases} \gamma_o = \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma_a^2 \\ \gamma_1 = \phi_1\gamma_o + \phi_2\gamma_1 \\ \gamma_2 = \phi_1\gamma_1 + \phi_2\gamma_o \end{cases}$$

or, if we are interested only in the autocorrelations, we may divide both sides of the last two equations of this set by γ_o to obtain

$$[25] \quad \begin{cases} \rho_1 = \phi_1 + \phi_2\rho_1 \\ \rho_2 = \phi_1\rho_1 + \phi_2 \end{cases}$$

These are called the Yule-Walker equations.

Autocorrelations of lag greater than 2 may be computed from the recursion relation

$$[26] \quad \rho_j = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} \quad (j > 2),$$

which results from

$$\gamma_j = E(\tilde{z}_t\tilde{z}_{t-j}) = E[(\phi_1\tilde{z}_{t-1} + \phi_2\tilde{z}_{t-2} + a_t)\tilde{z}_{t-j}]$$

$$= \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2}.$$

Note that Eq. [26] is formally the same as [23] without the disturbance term a_t.

For higher-order autoregressive processes, say AR(p), the model equation, in deviation-score form, is

$$[27] \quad \tilde{z}_t = \phi_1\tilde{z}_{t-1} + \phi_2\tilde{z}_{t-2} + \ldots + \phi_p\tilde{z}_{t-p} + a_t$$

where $\tilde{z}_t = z_t - \beta/(1-\phi_1-\phi_2-\ldots-\phi_p)$.

The Yule-Walker equations for computing $\rho_1, \rho_2 \ldots, \rho_p$ are p in number, and may best be displayed in matrix notation. They are:

$$[28] \quad \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{p-3} \\ \vdots & & & & \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}$$

The matrix on the right-hand side is symmetric with (i,j)- and (j,i)-elements equal to $\rho_{|i-j|}$. Thus, for instance, when p = 5, Eq. [28] reads

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \\ \rho_5 \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \rho_4 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \phi_3 \\ \phi_4 \\ \phi_5 \end{bmatrix}.$$

Autocorrelations of lag greater than p are given by the recursion relation

$$[29] \qquad \rho_j = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} + \ldots + \phi_p\rho_{j-p}, \quad (j > p)$$

i.e., an equation identical in form to the model equation [27] itself, except for the absence of a_t.

The expected value of z_t following a stationary AR(p) process is given by a simple extension of Eqs. [18] and [22], viz.:

$$[30] \qquad E(z_t) = \beta/(1-\phi_1-\phi_2-\ldots-\phi_p),$$

as was already anticipated when the deviation-score model equation was written.

Reciprocity between AR and MA Processes. What we saw in connection with the AR(1) model above exemplifies an interesting reciprocity that exists between autoregressive and moving-average processes: A finite autoregressive process is equivalent to an infinite moving-average process, while a finite moving-average process

is equivalent to an infinite autoregressive process. However, there is a slight asymmetry in the reciprocal relation.

Even for the simplest, finite autoregressive process AR(1) to be stationary, it was seen that ϕ_1 had to be less than one in absolute value. On the other hand, MA(1) (or any finite moving-average process, for that matter) is automatically stationary, as we saw earlier. Nevertheless, there is a sense in which the coefficient $-\theta_1$ in equation [7] for MA(1) needs to satisfy $|\theta_1| < 1$ in order for the process to be "reasonable." To show this, let us rewrite the equation for MA(1) in autoregressive form.

From Eq. [7], with t replaced by t-1, we get

$$a_{t-1} = z_{t-1} - L + \theta_1 a_{t-2},$$

which, substituted back in [7] yields

$$z_t = L + a_t - \theta_1(z_{t-1} - L + \theta_1 a_{t-2})$$

$$= L(1+\theta_1) - \theta_1 z_{t-1} + a_t - \theta_1^2 a_{t-2}.$$

Continuing in this manner, we eventually get

$$z_t = -\theta_1 z_{t-1} - \theta_1^2 z_{t-2} - \theta_1^3 z_{t-3} - \ldots + L(1+\theta_1+\theta_1^2+\ldots) + a_t.$$

Thus, even though MA(1) is known to be stationary, its rewriting in autoregressive form does not make sense unless $|\theta_1| < 1$. The right-hand side would "explode" if $|\theta_1| \geq 1$. Hence, we must require $|\theta_1| < 1$ for MA(1) even though no such condition was necessary for stationarity of MA(1) in its own right. This is called the invertibility condition for MA(1). Analogously, the requirement $|\phi_1| < 1$ is called the in-

vertibility condition for AR(1), even though in this case the condition is necessary also for an AR(1) process to be stationary.

For AR and MA processes of higher order, the invertibility conditions are more complicated, and we merely state them without derivation.

(a) For an AR(p) process to be stationary, the root of the characteristic equation

$$1 - \phi_1 - \phi_2 x^2 - \ldots - \phi_p x^p = 0$$

must lie outside the unit circle. [This anticipates that at least some of the roots will generally be complex. For any real root x_o, the requirement is simply that $|x_o| > 1$. Note that, for p = 1, this reduces to the earlier condition, $|\phi_1| < 1$. For then the characteristic equation is $1 - \phi_1 x = 0$, whose root is $x_o = 1/\phi_1$, so that $|x_o| > 1$ is equivalent to $|\phi_1| < 1$.]

(b) For a MA(q) process to be meaningfully expressible in autoregressive form, the roots of the characteristic equation

$$1 - \theta_1 - \theta_2^{\ 2} - \ldots \theta_q x^q = 0$$

must lie outside the unit circle.

THE MIXED MODEL: ARMA

Given the two basic models, AR(p) and MA(q), for stationary processes, it is a natural extension to form a combination of the two resulting in the ARMA (p,q) model (an autoregressive moving-average

model of order p,q), with the equation

$$[30] \quad z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots + \phi_p z_{t-p} + \beta + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

The advantage of making such a combination is implicit in the above discussion of the reciprocity between AR and MA processes. A finite AR process was shown to be equivalent to an infinite MA process, and vice versa. What is more to the point here are the equivalences in the opposite directions: an infinite MA process (or a finite one with a very large order q) may be expressible as an AR process of very small order, and, conversely, an AR process of very large order p may be expressible as an MA process of low order. Combining the two would, then, give us the best of two worlds, so to speak. Thus, a stationary process which cannot be expressed either by a pure MA or a pure AR model of reasonably low order may be expressible as a mixed ARMA (p,q) model with quite small orders p and q. The savings in the number of parameters to be estimated may be enormous.

The technicality of deriving the variance, autocovariances and autocorrelations for the ARMA(p,q) model is tedious, although in principle it involves no more than a combination of the procedures described above for MA(q) and AR(p) models separately. Since our main purpose here is simply to point out the advantage of sometimes considering the combined ARMA model, we shall not go into these derivations. We merely state the results for the simplest case, ARMA(1,1).

The model equation for ARMA(1,1) is

$$[31] \quad z_t = \phi_1 z_{t-1} + \beta + a_t - \theta_1 a_{t-1}.$$

It can be shown that

$$E(z_t) = \frac{\beta}{1 - \phi_1} \quad \text{[the same as for AR(1)]}$$

$$\gamma_o = \frac{1 + \theta_1^2 - 2\phi_1\theta_1}{1 - \phi_1^2}\,\sigma_a^2$$

$$\gamma_1 = \frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1 - \phi_1^2}\,\sigma_a^2$$

and $$\gamma_j = \phi_1\gamma_{j-1} \quad (j \geq 2).$$

Note that these results reduce to those for MA(1) when $\phi_1 = 0$, and to those for AR(1) when $\phi_1 = 0$. The autocorrelations are immediately obtainable by division: $\rho_j = \gamma_j/\gamma_o$, so we shall not list their formula here.

MODELS FOR NONSTATIONARY PROCESSES

Stationary time series are seldom "literally true" descriptions of processes encountered in practice, although they often provide good approximations. But sometimes--perhaps often in behavioral-science applications--they are not even adequate approximations, as when learning or growth is involved.

Fortunately, however, many nonstationary time-series observations that occur in real life exhibit what is known as homogeneous nonstationarity, by which is meant that even though the series moves about freely without centering around a fixed mean, its behavior is essentially similar throughout the course of time. When this is true, it often turns out that the series formed by the successive differences between adjacent observations,

[32] $w_t = z_t - z_{t-1}$,

is a stationary time series.

Sometimes, we may have to form second-order differences,

$$v_t = w_t - w_{t-1} = z_t - 2z_{t-1} + z_{t-2},$$

or even higher-order differences before stationarity is achieved. At any rate, the stationary models previously described for MA, AR and ARMA processes are usually found to be applicable to differences of suitable order d of observations following a nonstationary process. Thus, the most general model for nonstationary processes is one in which the d^{th} order differences constitute an ARMA(p,q) process. This is known as an integrated autoregressive moving-average process of order p, d, q and is symbolized ARIMA(p,d,q).[1]

---

[1]The qualifier "integrated" simply means that the terms of the original series $\{z_t\}$ are sums (of order d) of the d^{th} order differences which follow ARMA(p,q). For example, when d = 1,

$$\begin{aligned} z_t &= (z_t - z_{t-1}) + (z_{t-1} - z_{t-2}) + (z_{t-2} - z_{t-3}) + \cdots \\ &= w_t + w_{t-1} + w_{t-2} + \cdots \\ &= \sum_{i=0}^{\infty} w_{t-1}; \end{aligned}$$

Similarly, when d = 2, since w_t is itself the sum of present and all past v_t's, it follows that

$$z_t = \sum_{i=0}^{\infty} w_{t-1} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} v_{t-i-j},$$

a double sum of the second-order differences.

The equation for ARIMA(p,1,q), written in terms of w_t, is simply the ARMA(p,q) equation for w_t; i.e.,

$$[33] \quad w_t = \phi_1 w_{t-1} + \phi_2 w_{t-2} + \ldots + \phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}$$

Note, however, that there is one difference between this equation and the equation, [31], for the ARMA(1,1) process in z_t itself [which can be readily generalized to ARMA(p,q)], in that [33] does not contain the constant term β. Since the mean of ARMA(p,q) is the same as that of AR(p), as shown for ARMA(1,1) after Eq. [31], it follows from Eq. [30] that $E(w_t) = 0$. Thus the average of $z_t - z_{t-1}$ over a long period of time is approximately zero. For the original time series $\{z_t\}$, this implies that even though it does not center around a fixed mean, nor does it show a perpetual trend upward or downward. Technically, this is characterized by saying that z_t shows a stochastic trend on drift, but not a deterministic one. This is the situation usually treated in time-series analysis. In educational research where we usually expect learning to be taking place, it may well be that a deterministic trend should be incorporated. This can be done simply by adding a non-zero constant β to the right-hand side of Eq. [33]--although Box and Jenkins (1970, p. 93) advise against assuming a deterministic trend unless the data give clear evidence of its presence and form (linear, quadratic, etc.). Thus, the burden of the proof seems to be on including the constant term β to Eq. [33].

From the foregoing discussions, it is clear that nothing really new in the way of mathematical techniques is needed for handling homogeneous nonstationary time-series. We simply take differences of

sufficient order to achieve stationarity (as judged by methods discussed below), and apply the methods developed for AR(p), MA(q) or ARMA(p,q) processes, as the case may be. [Note that when $\theta_i = 0$, [33] reduces to the equation for an AR(p) process in w_t, while for $\phi_i = 0$ it reduces to that for a MA(q).]

There is, however, one new equation that it is sometimes convenient to have in dealing with ARIMA(p,d,q) processes, or their special cases, ARI(p,d) and IMA(d,q) processes. This is a rewriting of Eq. [33], or one of its more special instances, in terms of z_t in a form known as the (cumulative) random-shock form. We illustrate this for the simplest case, the IMA(1,1) process. The equation (replacing w_t by $z_t - z_{t-1}$) is

$$z_t - z_{t-1} = a_t - \phi_1 a_{t-1},$$

or

$$[34] \qquad z_t = z_{t-1} + a_t - \theta_1 a_{t-1},$$

which, it may be noted incidentally, formally resembles an ARMA(1,1) equation but nevertheless cannot be so construed, since the autoregressive coefficient is $\phi_1 = 1$ (cf. Eq. [31]), thus violating the stationarity condition $|\phi_1| < 1$.

Using [34] with t replaced by t - 1, we have

$$z_{t-1} = z_{t-2} + a_{t-1} - \theta_1 a_{t-2},$$

which may be substituted back in [34] to eliminate z_{t-1}:

$$z_t = (z_{t-2} + a_{t-1} - \theta_1 a_{t-2}) + a_t - \theta_1 a_{t-1}$$

$$= z_{t-2} + a_t - (1-\theta_1) a_{t-1} - \theta_2 a_{t-2}$$

Successively eliminating z_{t-2}, z_{t-3}, etc. in this manner, we eventually get

$$z_t = a_t + (1-\theta_1)(a_{t-1}+a_{t-2}+\ldots),$$

where the sum of the a_i's extends indefinitely into the past. It is convenient to break this sum down into two parts,

$$\sum_{i=-\infty}^{k} a_i \text{ and } \sum_{i=k+1}^{t-1} a_i,$$

where k is an arbitrary reference point. We may then write

$$z_t = (1-\theta_1) \sum_{i=-\infty}^{k} a_i + (1-\theta_1) \sum_{i=k+1}^{t-1} a_i + a_t,$$

or, upon denoting the first partial sum by L_k,

$$[35] \qquad z_t = L_k + (1-\theta_1) \sum_{i=k+1}^{t-1} a_i + a_t.$$

From the strict mathematical standpoint, the first partial sum $(1-\theta_1) \sum_{i=-\infty}^{k} a_i$ above may not even converge, and hence we have no right to denote this by L_k. However, from the practical standpoint we may reasonably assume that the disturbances beyond some remote time in the past should not affect the present observation, so that $a_j = 0$ for those remote time points. It is important to remember, however, that how far back is remote enough will depend on what the present time point t is. Thus, L_k is not strictly a constant, but depends in an indirect way on t. (This is what keeps Eq. [35] from representing a stationary process.) L_k may be interpreted as the "level" of the system at time point k.

IDENTIFYING THE PROCESS AND ESTIMATING ITS PARAMETERS

The foregoing concludes our discussion--necessarily incomplete because our aim was to keep it as elementary as possible--of the various models, stationary and nonstationary, for time-series observations. We now come to the practical question: given a set of time-series data, how do we identify which of the several models is appropriate, and how do we estimate the parameters of the selected model?

It is at this point that we part company from the traditional procedures of time-series analysis and propose alternative methods which we believe to be better adapted to data from longitudinal studies. But first we must outline the traditional methods and point out the difficulties in applying them to longitudinal data.

Traditional Procedures

Since, as indicated earlier, nonstationary processes of the homogeneous variety are adequately modeled by stationary processes--AR(p), MA(q) and ARMA(p,q)--in the differences of suitable order, we shall confine our discussions primarily to stationary models.

The behavior of the sample counterparts of the autocorrelations of various lags is, as mentioned earlier, the key to identifying the appropriate model for a given set of time-series data. We therefore first indicate how the sample autocorrelations r_j corresponding to the theoretical parameters ρ_j have traditionally been defined and computed.

The Sample Autocorrelation r_j. Historically, there have been several alternative definitions proposed for r_j, but the one currently favored is as follows:

Given an observed series of data $z_1, z_2, \ldots, z_T$ at T time

points, we compute the sample variance c_o and sample autocovariances of lag j, c_j, as

$$[36] \quad \begin{cases} c_o = \frac{1}{T} \sum_{t=1}^{T} (z_t - \bar{z})^2 \\ c_j = \frac{1}{T} \sum_{t=1}^{T-j} (z_t - \bar{z})(z_{t+j} - \bar{z}), \quad j = 1, 2, \ldots \end{cases}$$

Where $\bar{z}$ is the sample mean

$$\bar{z} = \frac{1}{T} \sum_{t=1}^{T} z_t$$

Based on the sample variance and autocovariances, the sample autocorrelation of lag j is defined as

$$[37] \quad r_j = c_j/c_o, \quad j = 1, 2, \ldots$$

Once the sample autocorrelations have been computed, and possibly plotted against j for visual inspection of their behavior, we check to see if the trend with j corresponds approximately to the trend exhibited by the theoretical autocovariances ρ_j for any of the models MA(q), AR(p) or ARMA(p,q).

Identification of an MA(q) Process. If an observed time series conforms (approximately) to an MA(q) process, this fact is readily discernible by inspection of the trend of r_j with j. As stated in the discussion preceding Eq. [13], the theoretical autocorrelations for an MA(q) process are non-zero for lags up to and including q, and then abruptly drop to zero. If the sample autocorrelations show this sort of trend with j, we may safely conclude that a moving-average model adequately fits the data, with order equal to the last j for which r_j is substantially non-zero. For instance, if r_1 alone is of considerable

magnitude while r_2, r_3, ... are essentially zero, we conclude that the data are adequately modeled by an MA(1) process; if r_1 and r_2 are substantially non-zero and the rest (r_3, r_4, ...) are of trivial magnitude, we conclude that MA(2) offers an adequate fit.

There are significance tests available for judging when a sample autocorrelation is "substantially non-zero" and when it is "essentially zero" within sampling error, but we shall not discuss these in this brief outline. The interested reader may refer to Box and Jenkins (1970, pp. 177-78), Glass et al. (1975, pp. 97-98) or Nelson (1973, pp. 71-72).

Identification of an AR(p) Process. Except when the observed data sequence is adequately modeled by an AR(1) process, the identification of the appropriate order p of an autoregressive process fitting the data is much more difficult than in the moving-average case.

As shown in Eq. [19], the theoretical autocorrelations for AR(1) exhibit an exponential decay with increasing lag j. If the sample autocorrelations more or less follow this pattern--i.e., decreasing geometrically with lag j but not suddenly dropping to a near-zero value from a certain j on--we are fairly safe in concluding that an AR(1) model will fit the data adequately.

When the above happy circumstance does not prevail (and the fitting of a moving-average model has already been ruled out), things get much more complicated. Inspection of the behavior of autocorrelations alone will not suffice, and we must examine what are known as partial autocorrelations.

The basic rationale hinges on the relation between the autoregressive cofficients ϕ_i and the autocorrelations ρ_j specified by

the Yule-Walker equations (see Eqs. [25] and [23]). We know that for AR(p) the coefficients ϕ_i for $i > p$ must vanish; the Yule-Walker equations enable us successively to estimate the ϕ_i's by using the sample autocorrelations r_j in place of the theoretical ρ_j, and hence to detect for what i ϕ_i first becomes essentially zero.

Assuming that AR(1) has been ruled out by the r_j's not decaying approximately exponentially, we wish to check if an autoregressive process of order 2 or greater will fit the data. We replace the ρ_1 and ρ_2 in the Yule-Walker equations [25] by their sample estimates r_1 and r_2, thus:

$$r_1 = \hat{\phi}_1 + \hat{\phi}_2 r_1$$

$$r_2 = \hat{\phi}_1 r_1 + \hat{\phi}_2$$

where we have also replaced the ϕ_i by $\hat{\phi}_i$ to signify that we are solving for estimates of ϕ_i. If the solution for $\hat{\phi}_2$ differs significantly from zero, we conclude that the order of the AR process is at least 2, and proceed to the next step of checking if the order is 3 or greater. That is, we solve the Yule-Walker equations with $p = 3$ for $\hat{\phi}_3$:

$$r_1 = \hat{\phi}_1 + \hat{\phi}_2 r_1 + \hat{\phi}_3 r_2$$

$$r_2 = \hat{\phi}_1 r_1 + \hat{\phi}_2 + \hat{\phi}_3 r_1$$

$$r_3 = \hat{\phi}_1 r_2 + \hat{\phi}_2 r_1 + \hat{\phi}_3$$

If the solution for $\hat{\phi}_3$ is significantly different from zero, we proceed to $p = 4$, and so on. Eventually, we will come to a $\hat{\phi}_{p^*}$ that does not differ significantly from zero, and we then conclude that AR(p*-1)

offers an adequate fit of the data (assuming, of course, that a pure AR model is appropriate in the first place).

Although we have not used the term "partial autocorrelation" in the above discussion, this is the name given to the $\hat{\phi}_j$ solved from the Yule-Walker equations with p = j, and it is conventionally denoted $\hat{\phi}_{jj}$, the sample partial autocorrelation of order j. This may be computed by Cramer's rule [see Box and Jenkins (1970, p. 64)] without having to solve the Yule-Walker equations in their entirely.

The significance test for $\hat{\phi}_{jj}$ is quite simple, for it has been shown by Quenouille (1949) that the approximate standard error of $\hat{\phi}_{jj}$ is $1/\sqrt{T}$ when $\phi_j = 0$. Thus, we have merely to multiply the computed value of $\hat{\phi}_{jj}$ by $\sqrt{T}$ (T being the number of data points) and refer to a normal-curve table.

Identification of an ARMA(p,q) Process. If a pure MA model has been ruled out, and the partial autocorrelation $\hat{\phi}_{jj}$ does not drop to nonsignificance for a long time (i.e., until j exceeds 3 or 4, say), then we must suspect that a mixed ARMA model may offer a better fit to the data with lower orders p and q. (See discussion in section on the mixed model.) Unfortunately, the identification of the orders of an ARMA process is even more complicated a task than identifying the order of a pure AR process.

About all we can say is that, when both sample autocorrelations and sample partial autocorrelations decline gradually rather than dropping abruptly to near-zero, a mixed process is indicated. As a working rule, it may be said that it is worth considering an ARMA model only if both orders are no greater than 2; i.e., ARMA(1,1), ARMA(1,2), ARMA(2,1) and ARMA(2,2) are the only models that should be

entertained seriously after pure MA and pure AR models of reasonably low order have been ruled out. Beyond that, it is probably more fruitful to postulate a nonstationary model.

Summary of Process-Identification Rules. We may summarize the foregoing procedures for identifying an appropriate stationary time-series process for modeling a sequence of observed data in the form of a table listing the rules-of-thumb. It should always be borne in mind that the orders should be relatively low (no higher than 3, perhaps, for pure MA and AR models, and no higher than (2,2) for the mixed ARMA model) for us to consider a stationary model seriously.

Table 1. Behaviors of autocorrelations and partial autocorrelations in various processes

Process	Autocorrelations	Partial Autocorrelations
MA(q)	Non-zero for lags 1 through q; then abruptly drop to 0	(Taper off; but not necessary to check)
AR(1)	Taper off exponentially	Only $\hat{\phi}_{11} \neq 0$
AR(p), p>1	Taper off according to $\rho_1 = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} + \cdots + \phi_p\rho_{j-p}$	$\hat{\phi}_{11}, \hat{\phi}_{22}, \ldots, \hat{\phi}_{pp} \neq 0$ $\hat{\phi}_{jj} = 0$ for $j > p$
ARMA(p,q)	Irregular pattern for lags 1 through q; then taper off according to $\rho_1 = \phi_1\rho_{j-1} + \cdots + \phi_p\rho_{j-p}$	Taper off

Recognizing Nonstationarity. If the sample autocorrelations taper off very gradually over a long stretch of lags, we have prima facie evidence that a nonstationary process is indicated. MA(q) is certainly ruled out immediately, and even if an AR(p) process should be appropriate,

it is likely that the order p will be quite large. If the checking of the first few partial autocorrelations (through $\hat{\phi}_{33}$, say) confirms this by their being of considerable magnitude ($\hat{\phi}_{33}\sqrt{T} > 2$, say), AR(p) may be ruled out for practical purposes. ARMA(p,q) should probably not be considered unless the partial autocorrelations taper off rather rapidly.

Besides the above considerations, it is always a good idea to make a plot of z_t against t to get a visual impression of the lack of stationarity—although one should not rely entirely on visual impressions. At any rate, if the data are from an area in educational research such that learning is expected to take place within the period of observation, it is more likely than not that the series will display nonstationarity, as mentioned several times earlier. Such being the case, it is probably wise not to expend a large amount of time and effort in seeking to make a Procrustean fit of the data series to some stationary model. Rather, one should adopt the standpoint that nonstationarity exists unless clear and quick (i.e., with low orders p, q, or (p, q) for the model) evidence is available to the contrary.

Once we decide on a nonstationary model, we form the first differences $w_t = z_t - z_{t-1}$ and treat the series $w_1, w_2, \ldots, w_{T-1}$ just as we did the original observed series. That is, we determine the autocorrelations and (if necessary) the partial autocorrelations of this new series and check if an MA, AR or ARMA model of reasonably low order will adequately fit the data. If so, we conclude that the original series $z_1, z_2, \ldots, z_T$ is adequately modeled by an IMA(1,q), ARI(p,1) or ARIMA(p,1,q) process. If not, it must be concluded that even the series of the first-order differences exhibit nonstationarity. We then take the second differences $v_t = w_t - w_{t-1}$ and repeat the entire search

procedure with the series $v_1, v_2, \ldots, v_{T-2}$. Fortunately, experience shows that we rarely, if ever, need to go beyond the second-order differences to achieve stationarity.

Estimation of Parameters. Once an appropriate model has been identified, we may estimate the parameters of the model by using the sample autocorrelations. In outline, what we do is to substitute the sample autocorrelation values for the theoretical autocorrelations in the equations relating the latter to the basic parameters, and solve the resulting equation(s).

For pure AR processes, the procedure is straightforward. In particular, for AR(1) we need only take the $j = 1$ instance of Eq. [19] to get

$$\hat{\phi}_1 = r_1.$$

[For somewhat greater accuracy of estimation, we might take the first few instances,

$$\hat{\phi}_1 = r_1, \ \hat{\phi}_1^2 = r_2, \ \hat{\phi}_1^3 = r_3 \text{ (say)},$$

and get a least-squares estimate for $\ln\phi_1$.]

For AR processes of higher order, we may substitute the values of r_j for the corresponding ρ_j's in the Yule-Walker equations (see Eq. [28]) and solve the set of linear equations for the $\hat{\phi}_i$. Thus, for example, for $p = 3$ (beyond which we would seldom wish to go), the indicated substitutions in Eq. [28] yield

$$\begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix} = \begin{bmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_1 \\ r_2 & r_1 & 1 \end{bmatrix} \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \hat{\phi}_3 \end{bmatrix},$$

from which we immediately get

$$\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \hat{\phi}_3 \end{bmatrix} = \begin{bmatrix} 1 & r_1 & r_2 \\ r_1 & 1 & r_1 \\ r_2 & r_1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \end{bmatrix}$$

For MA and ARMA processes, the procedures are somewhat more complicated because the relations between the autocorrelations and the basic parameters are nonlinear. Thus, for AM(1), we have, upon substitution of r_1 for ρ_1 in Eq. [12],

$$r_1 = \frac{-\hat{\theta}_1}{1 + \hat{\theta}_1^2},$$

which is a quadratic in $\hat{\phi}_1$ with two solutions

$$[38] \qquad \hat{\phi}_1 = \frac{-1 \pm \sqrt{1 - 4r_1^2}}{2r_1}$$

It is easily verified that the two solutions are reciprocals of each other. Hence, just one of them must satisfy the invertibility condition, $|\theta_1| < 1$. This is the one we take as our estimate for θ_1.

The equations become much more complicated for MA(2). Substitution of r_1 and r_2 for ρ_1 and ρ_2, respectively, in Eqs. [13] yields

$$r_1 = \frac{-\hat{\theta}_1 + \hat{\theta}_1\hat{\theta}_2}{1 + \hat{\theta}_1^2 + \hat{\theta}_2^2}$$

and

$$r_2 = \frac{-\hat{\theta}_2}{1 + \hat{\theta}_1^2 + \hat{\theta}_2^2}.$$

Simultaneous iterative solution of these two equations for $\hat{\theta}_1$ and $\hat{\theta}_2$ has been the usual approach. The present writer has found, however,

after some algebraic manipulations, that we may equivalently solve for $\hat{\theta}_1$ from the equation

$$[39] \qquad r_1 = \hat{\theta}_1 r_2 \left[\frac{2r_2}{1 - \sqrt{1 - 4r_2^2(1+\hat{\theta}_1^2)}} + 1 \right]$$

and then obtain $\hat{\theta}_2$ from

$$[40] \qquad \hat{\theta}_2 = \frac{-1 + \sqrt{1 - 4r_2^2(1+\hat{\theta}_1^2)}}{2r_2}.$$

This simplifies the solution in that iteration (by, e.g., the Gauss-Newton Method) needs to be carried out only on Eq. [39], with one unknown, $\hat{\theta}_1$. The closed expression [40] then yields $\hat{\theta}_2$. Also, the satisfaction of the invertibility condition, that the roots of

$$1 - \theta_1 x - \theta_2 x^2 = 0$$

must lie outside the unit circle, is built into Eqs. [39] and [40] provided only that we take the solution of [39] with $|\theta_1| < 1$.

The procedures for MA processes of order q greater than 2 are, needless to say, even more complicated. Simultaneous iterative solution, by the Gauss-Newton method, of the system of nonlinear equations (generalized from Eqs. [13])

$$r_1 = \frac{-\hat{\theta}_1 + \hat{\theta}_1\hat{\theta}_2 + \ldots + \hat{\theta}_{q-1}\hat{\theta}_q}{1 + \hat{\theta}_1^2 + \hat{\theta}_2^2 + \ldots + \hat{\theta}_q^2}$$

$$r_2 = \frac{-\hat{\theta}_2 + \hat{\theta}_1\hat{\theta}_3 + \ldots + \hat{\theta}_{q-2}\hat{\theta}_q}{1 + \hat{\theta}_1^2 + \ldots + \hat{\theta}_q^2}$$

$$\vdots$$

$$r_q = \frac{-\hat{\theta}_q}{1 + \hat{\theta}_1^2 + \ldots + \hat{\theta}_q^2}$$

is about all that can be hoped for.

For mixed processes ARMA(p,q), the estimation procedure is too complicated to expound here, except in gross outline, for all but the simplest case, ARMA(1,1). In the general case, recursion relations (similar to those for the pure AR process--cf. Eq. [29]) exist between $\rho_{q+1}, \rho_{q+2}, \ldots, \rho_{q+p}$ and the autocorrelations of lag q or less. From these relations, estimates $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p$ for $\phi_1, \phi_2, \ldots, \phi_p$ can be computed. Then, utilizing the relations between $\rho_1, \rho_2, \ldots, \rho_q$ and the $\hat{\phi}_i$'s and the θ_j's, we may solve for $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_q$ by substituting r_j for ρ_j and $\hat{\phi}_i$ for ϕ_i.

For the simplest case, ARMA(1,1), the details are as follows: from the fourth equation after [31], letting j = 2, we get

$$\rho_2 = \phi_1 \rho_1,$$

from which (after replacing ρ_1 and ρ_2 by r_1 and r_2, respectively)

$$\hat{\phi}_1 = \frac{r_2}{r_1}$$

Then, from two other equations following [31],

$$r_1 = \frac{(1-\hat{\phi}_1\hat{\theta}_1)(\hat{\phi}_1-\hat{\theta}_1)}{1 - \hat{\theta}_1^2 - 2\hat{\phi}_1\hat{\theta}_1},$$

in which $\hat{\phi}_1 = r_2/r_1$ may be substituted, and the resulting equation solved for $\hat{\theta}_1$. (Alternative solutions for $\hat{\theta}_1$ will again be obtained, among which the one satisfying the invertibility condition is chosen.)

Now, all the parameter estimates described above are, in the traditional approach, taken to be "preliminary estiamtes" only. After these are obtained, it is customary to use maximum-likelihood methods [which in this case turns out to be equivalent to minimizing

the sum of squares for lack of fit, $\sum(z_t - \hat{z}_t)^2$], employing the preliminary estimates as the starting values for the complicated iterative procedures that have to be used. We shall not discuss this refinement here because, as will be argued later, it seems to be unnecessary when we use the alternative estimation procedure for longitudinal data to be proposed below.

Difficulties with the Traditional Procedures When Applied to Longitudinal Data

The reader will have noticed that, throughout the foregoing discussions, it was assumed that there is but one observation z_t at each point in time. This is necessarily the case in economic or demographic applications of time-series analysis, where, for example, the consumer price index or the unemployment rate in successive years (or quarters or months) constitute the observations z_t.

For longitudinal data from an intact group being observed at a series of time points 1, 2, ..., T, however, there are N observations $z_{1t}, z_{2t}, \ldots, z_{Nt}$ at each time point t(= 1, 2, ..., T). That is to say, instead of a vector of data

$$\underset{\sim}{z}' = [z_1, z_2, \ldots, z_T]$$

we have an N x T data matrix (where N is the group size)

$$\underset{\sim}{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1T} \\ z_{21} & z_{22} & \cdots & z_{2T} \\ \vdots & & & \\ z_{N1} & z_{N2} & \cdots & z_{NT} \end{bmatrix}$$

as our input series. True, the input data can always be condensed into a row vector by considering only the group mean $\bar{z}_{\cdot t}$ at each time point as our input (as we would be forced to do in order to apply the traditional procedures as they stand). But this obviously does violence to the data, and throws away a lot of potential information contained in the separate rows of the data matrix $\underset{\sim}{Z}$--or, otherwise stated, ignores the correlatedness of the T observations across each row. To draw an analogy with analysis of variance, it is akin to using a randomized-groups design when a repeated-measures design is the correct model to use. This, then, is the major difficulty the present writer sees with the traditional procedures of time-series analysis when it is to be applied to data from longitudinal studies.

Another difficulty with the traditional procedures is that it requires a large number of time points T at which the (single) observations are taken. Box and Jenkins (1970) assert that ". . .to obtain a useful estimate of the autocorrelation function, we would need at least fifty observations [i.e., $T \geq 50$] . . ." (p. 33). It is obviously too much to expect so many time points of observation in a longitudinal study, unless the unit of time is as short as a day, or at most a month. But normally, in educational research, we would not be interested in such short time units. An year, a semester, or at least a quarter, would more likely be the interval between successive observations. Thus, the number of time points will usually be in the range 5-20 instead of the minimum of 50 recommended by Box and Jenkins.

It was precisely in an attempt to resolve the foregoing difficulties that the present research was undertaken. It seemed intuitively clear that having, say, N = 30 observations at each of T = 10 time points

should, in some sense, yield nearly as much information (although of course not just as much) as having 300 time points, each with one observation. To look only at the 10 mean observations, $\bar{z}_{\cdot 1}$, $\bar{z}_{\cdot 2}$, ..., $\bar{z}_{\cdot 10}$, seems to be a gross waste of data.

It should be mentioned that Glass, Willson and Gottman (1975) have implicitly addressed themselves to this problem by discussing (in their first chapter) the distinction between unit-repetitive and unit-replicative designs. The former refers to the case when an intact group is "observed at several successive points in time"--i.e., the genuine longitudinal study. The latter refers to the case when samples from the same conceptual population (e.g., the population of car drivers in a certain state in successive years)--but one which does not comprise the same set of individuals over time--are observed at successive time points. Although they acknowledge the importance of both designs and even point out that use of the unit-replicative design may sometimes be invalid (as when a change of composition of the population occurs from before to after an intervention), they opt to deal, in their subsequent chapters, solely with procedures that are adapted to the unit--replicative design. Thus, the substantive examples they present concern such phenomena as the "percentage of students in Ireland who passed the intermediate and senior level examinations of the years 1879-1924," "the number of traffic fatalities per 100,000,000 driver miles in the state of New York for the 100 months from January 1951 to April 1960," and the "petition for reconciliation rate. . .in German states. . .prior to and for fourteen years after institution of the new Civil Code of the German Empire on January 1, 1900."

One cannot, of course, fault the authors for their particular

choice of design (the unit-replicative design) on which to concentrate in their book. But the fact remains that--valuable as their pioneering efforts in bringing time-series analysis to the attention of educational researchers have been--they have not specifically considered the case of longitudinal studies, despite their frequent mention of this phrase.

Estimation Procedures Geared to Longitudinal Studies

After a number of trial-and-error attempts at developing model-identification and parameter-estimation procedures especially geared to the application of time-series analysis to longitudinal data (i.e., the unit-repetitive design, in Glass et al.'s terminology) the only viable procedure discovered to date was the "obvious" one of utilizing the ordinary sample correlation matrix based on the data matrix $\underset{\sim}{Z}$. This is "obvious" only in retrospect, however, since the use of the correlation matrix for estimating the autocorrelations carries with it the assumption that the observations $\{z_{it}\}$ for every individual follows the same stochastic process with the same parameters, which is clearly a strong assumption. (More will be said about this later.)

Once it is decided to use the T x T sample correlation matrix $\underset{\sim}{R}$ based on the data matrix $\underset{\sim}{Z}$ for estimating the autocorrelations, the details of how to do so remain to be developed. The simplest way is to treat the average of the correlations r_{ik} with subscripts such that $i - k = j$ as an estimate of ρ_j, the theoretical autocorrelation of lag j. That is, the mean of the correlations along the line parallel and adjacent to the main diagonal of $\underset{\sim}{R}$ is used as an estimate of ρ_1; the mean of the correlations along the next line to the left and below this is

outlined above for the traditional approach. Details are best relegated to a couple of numerical examples, one using real data and the other based on simulated data. The functions of these numerical examples are twofold: first, to provide some evidence of the validity of the proposed parameter-estimation (and hence also of the model-identification) procedure; and second, to illustrate the method for detecting and significance-testing an intervention effect as developed by Glass et al. (1975). The latter is not expounded here except in the context of the numerical examples for two reasons. First, the present writer is unable to improve upon (i.e., expound in a more elementary fashion than) the original exposition by Glass and his coworkers. Second, the writer believes that there must be a way more consonant with longitudinal data for detecting and testing intervention effects, but has so far been unable to discover one. Hence, the method developed by Glass et al. is here used as a "stop-gap" measure rather than something the writer would advocate in earnest for longitudinal studies. (This is not to detract from its merits as a method used in conjunction with unit-replicative as against unit-repetitive designs.)

NUMERICAL EXAMPLES[3]

Our first example is based on data from a study investigating

---

[3]All computations were done by K. Tatsuoka on the PLATO system at the Computer-based Education Research Laboratory, University of Illinois at Urbana-Champaign.

outlined above for the traditional approach. Details are best relegated to a couple of numerical examples, one using real data and the other based on simulated data. The functions of these numerical examples are twofold: first, to provide some evidence of the validity of the proposed parameter-estimation (and hence also of the model-identification) procedure; and second, to illustrate the method for detecting and significance-testing an intervention effect as developed by Glass et al. (1975). The latter is not expounded here except in the context of the numerical examples for two reasons. First, the present writer is unable to improve upon (i.e., expound in a more elementary fashion than) the original exposition by Glass and his coworkers. Second, the writer believes that there must be a way more consonant with longitudinal data for detecting and testing intervention effects, but has so far been unable to discover one. Hence, the method developed by Glass et al. is here used as a "stop-gap" measure rather than something the writer would advocate in earnest for longitudinal studies. (This is not to detract from its merits as a method used in conjunction with unit-replicative as against unit-repetitive designs.)

NUMERICAL EXAMPLES[3]

Our first example is based on data from a study investigating

---

[3]All computations were done by K. Tatsuoka on the PLATO system at the Computer-based Education Research Laboratory, University of Illinois at Urbana-Champaign.

possible learning (or practice) effects in completing cloze passages.[4] Fifty-two fifth grade pupils were given three cloze passages (one on sports, one on music and one "miscellaneous"--all passages being taken from a children's encyclopedia) to complete on each of 16 consecutive school days. The maximum possible score was 30 (10 for each passage). Complete data were abailable for 45 of the 52 subjects, so our input data matrix Z is of order 45 x 16. The column means--i.e., the group means for the 16 days--were as shown below, in Figure 1 shows their plot. No discernible learning effect is present.

$\bar{z}_{.1}$ - $\bar{z}_{.8}$	13.56	12.47	13.11	11.60	17.07	14.13	12.00	15.51
$\bar{z}_{.9}$ - $\bar{z}_{.16}$	16.69	13.31	13.47	10.00	13.13	12.60	12.22	12.47

The correlation matrix based on the data matrix $\underset{\sim}{Z}$ is shown in Table 2, along with the estimated correlations of lags 1 through 15, calculated in accordance with Eq. [41]. It is seen that the $\bar{r}_j$'s decline irregularly and very gradually over the entire span of 15 lags, which is a sign that nonstationarity may be present. (This view is corroborated by the visual impression provided by Figure 1.) To make sure that an autoregressive process of order 2 or 3 will not offer an adequate fit, however, let us compute the partial autocorrelation coefficients $\hat{\phi}_{33}$ and $\hat{\phi}_{44}$. The Yule-Walker equations for p = 3, with

---

[4] The study was conducted by a graduate student, Gregory Bell, under the supervision of our colleague Steven Asher. We are greatly indebted to Steve for making the data available to us.

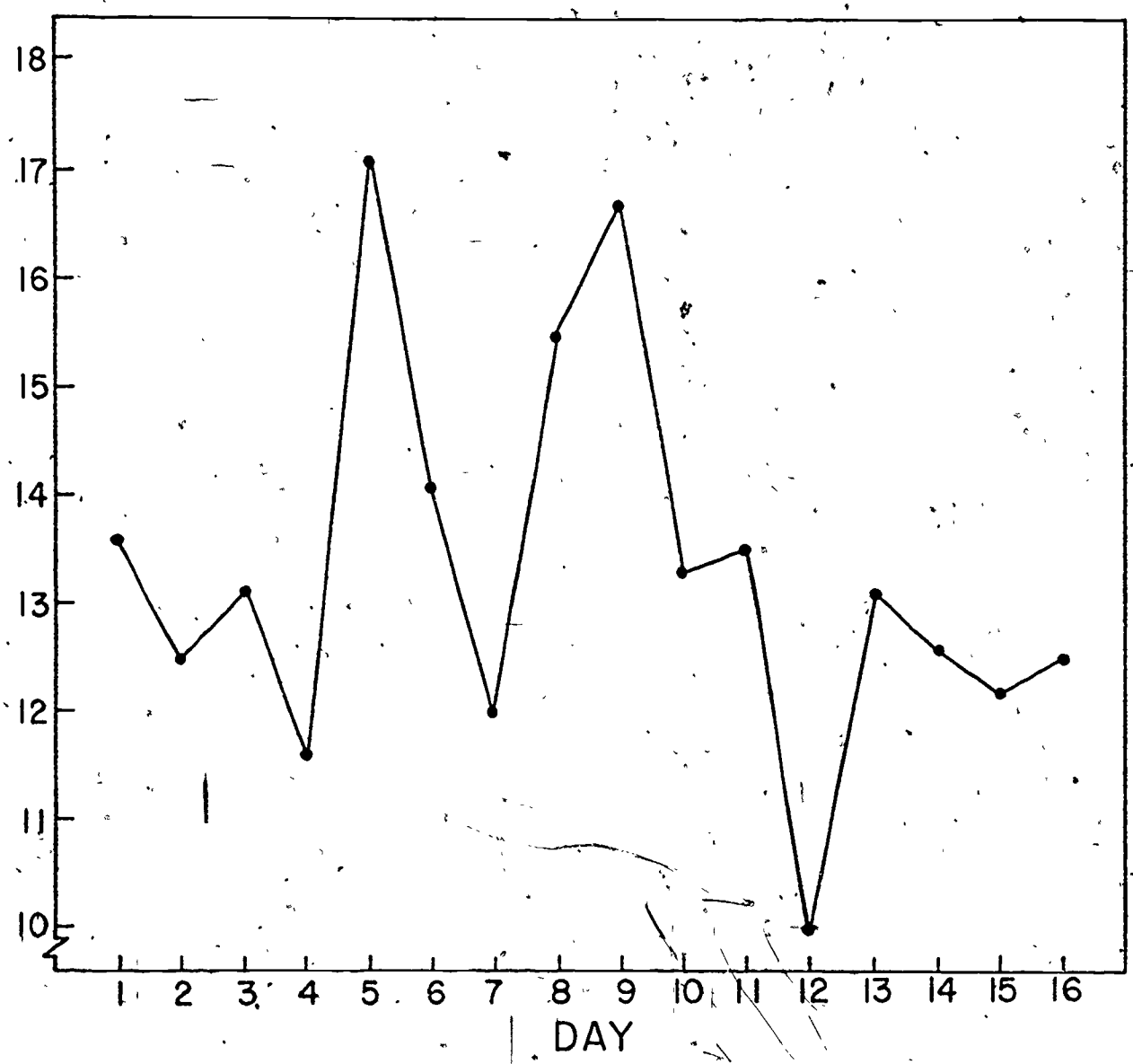


Figure 1. Mean cloze scores for a group of 45 fifth-graders on 16 consecutive school days.

Table 2. Correlation Matrix for scores earned by 45 fifth-graders on three cloze passages on 16 consecutive school days

1.00															
.65	1.00														
.59	.66	1.00													
.63	.67	.68	1.00												
.42	.55	.50	.58	1.00											
.39	.38	.61	.53	.58	1.00										
.68	.61	.72	.67	.44	.60	1.00									
.58	.56	.64	.66	.55	.64	.68	1.00								
.56	.50	.63	.61	.59	.57	.61	.81	1.00							
.60	.62	.68	.69	.69	.63	.78	.78	.79	1.00						
.53	.46	.54	.58	.46	.56	.65	.64	.61	.72	1.00					
.60	.66	.71	.59	.53	.59	.71	.61	.59	.65	.58	1.00				
.59	.67	.60	.65	.49	.51	.65	.65	.64	.72	.62	.64	1.00			
.48	.54	.59	.59	.63	.53	.65	.65	.67	.72	.69	.70	.64	1.00		
.49	.56	.52	.61	.43	.54	.61	.46	.46	.65	.43	.46	.53	.37	1.00	
.42	.33	.44	.38	.03	.23	.49	.37	.20	.42	.51	.36	.40	.20	.33	1.00

AVERAGE LAGGED CORRELATIONS

Lag:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\bar{r}_j$:	.621	.576	.604	.566	.609	.580	.520	.540	.584	.508	.500	.510	.492	.409	.417

the ρ_j replaced by $\bar{r}_j$ are, from Eq. [28],

$$\begin{bmatrix} 1 & .621 & .576 \\ .621 & 1 & .621 \\ .576 & .621 & 1 \end{bmatrix} \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \hat{\phi}_3 \end{bmatrix} = \begin{bmatrix} .621 \\ .576 \\ .604 \end{bmatrix}$$

since we are interested only in the value of $\hat{\phi}_3$ (which is the same as $\hat{\phi}_{33}$), we need not solve the entire system of equations for $\hat{\phi}_1$, $\hat{\phi}_2$ and $\hat{\phi}_3$. Using Cramer's rule, we have

$$\hat{\phi}_{33} = \frac{\begin{vmatrix} 1 & .621 & .621 \\ .621 & 1 & .576 \\ .576 & .621 & .604 \end{vmatrix}}{\begin{vmatrix} 1 & .621 & .576 \\ .621 & 1 & .621 \\ .576 & .621 & 1 \end{vmatrix}} = \frac{.1012}{.3412} = .297$$

Although this value is judged insignificant by the traditional significance test, for

$$\hat{\phi}_{33}\sqrt{T} = (.297)(4) = 1.19,$$

it should be borne in mind that the significance test is customarily used in conjunction with fairly large T ($\geq$ 50, say). For T as small as 16, it would require a $\hat{\phi}_{33}$ value of about .50 before it is judged significant. In situations like this, one should not rely heavily on significance tests. Regardless of its statistical insignificance, the value .297 is certainly a non-negligible one by any standard. If we were to adopt an AR model, we would certainly not be inclined to ignore

the third term with coefficient .297. Thus, the order of the presumed AR process will be at least 3.

Similarly, by solving the Yule-Walker equations with p = 4 for $\hat{\phi}_4$, we get $\hat{\phi}_{44}$ = .144, which is still not close enough to zero to be negligible. Thus, if we were to try to fit an AR model to the original data, we would need the order to be at least 4.

At this point, both common sense and the principle of parsimony would suggest that, instead of continuing to try to find a stationary model to fit the original data, it would be more strategic to go to the first differences, $w_t = z_t - z_{t-1}$. The new "data matrix" $\underset{\sim}{W}$ is now of order 45 x 15.

Table 3 shows the 15 x 15 correlation matrix of the w_t's, and the estimated autocorrelations of lags 1 through 14, again computed in accordance with Eq [41]. It is seen that $\bar{r}_j$ drops abrubtly to a near-zero value for j = 2, although there are a few, sporadic values that are not quite so small at larger lags. (The value -.215 for $\bar{r}_{14}$ may be discounted, since it is based on just one correlation value, $r_{15,1}$.)

Thus, it seems legitimate to entertain the MA(1) model for the sequence of first-order differences (which implies that the original series follows an IMA(1,1) process). I.e., we assume that

$$w_t = a_t - \theta_1 a_{t-1}.$$

The next step is to estimate θ_1 by means of Eq. [12] with ρ_1 replaced by $\bar{r}_1$. As we saw earlier, this equation has the solutions

$$\hat{\theta}_1 = \frac{-1 \pm \sqrt{1 - 4\bar{r}_1^2}}{2\bar{r}_1}$$

Table 3. Correlation matrix: based on first difference scores

1.00													
-.54	1.00												
.01	-.42	1.00											
.10	-.06	-.34	1.00										
-.26	.35	-.24	-.56	1.00									
.02	-.19	.08	-.34	-.19	1.00								
.00	-.01	.04	.20	-.09	-.63	1.00							
-.04	.10	-.01	.14	-.23	.06	-.13	1.00						
.19	-.17	.08	.06	-.11	.21	-.40	-.49	1.00					
-.19	.08	-.02	-.19	.30	-.13	.07	-.21	-.24	1.00				
.21	-.08	-.18	.06	-.05	.05	-.14	.02	-.07	-.47	1.00			
-.04	-.11	.21	-.11	-.01	-.03	.21	-.03	.03	-.01	-.58	1.00		
-.04	.17	-.07	.27	-.19	-.02	-.01	.08	-.05	.09	.00	-.54	1.00	
.03	-.10	.11	-.23	.21	-.02	-.15	-.09	.20	-.23	.01	.16	-.54	1.00
-.22	.20	-.22	-.13	.17	.08	-.15	-.29	-.04	.40	-.16	.00	-.01	-.49	1.00

AVERAGE LAGGED CORRELATIONS

Lag:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\bar{r}_j$:	-.440	-.093	.077	-.102	.130	-.026	-.080	-.013	.083	-.007	.030	-.009	.113	-.215

given by Eq. [38]. Substituting $\bar{r}_1 = -.440$ in this equation, we get

$$\hat{\theta}_1 = .5966 \text{ or } 1.6761,$$

of which the one with absolute value less than unity, viz., $\hat{\theta}_1 = .5966$ is the one we need.

Having obtained this estimate, how can we tell whether it is a "good" one? Unlike in the case of a deterministic model (such as a regression equation), we cannot verify the goodness of fit by computing estimated scores from the model equation and comparing (or correlating) them with the observed scores, for the model equation contains the unobservable random variable a_t. There are some complicated and indirect methods for checking the adequacy of the chosen model and estimated parameter(s). (See, e.g., Nelson, 1973, Section 5-11.) In our numerical example it was decided, after various considerations, to use the following approach, which seemed simpler than existing techniques and adequate for our purpose. (It also has the advantage of illustrating, in its simplest form, the general method developed by Glass et al., 1975, for estimating and testing intervention effects.)

Suppose we imagine a fictitious intervention between days 8 and 9 such that leads to an immediate elevation of the "level" of the system by a specified number of units, say 5 points. The modified plot of group means, with all points from day 9 on moved upwards by 5 units from their original positions in Figure 1, is shown in Figure 2. Of course, this constant elevation of scores will not affect the correlations among either the original z_t's or the first differences w_t. Hence, the estimate of θ_1 will remain unchanged. We may then ask the following question: Using the previously estimated $\hat{\theta}_1 = .5966$ in the technique

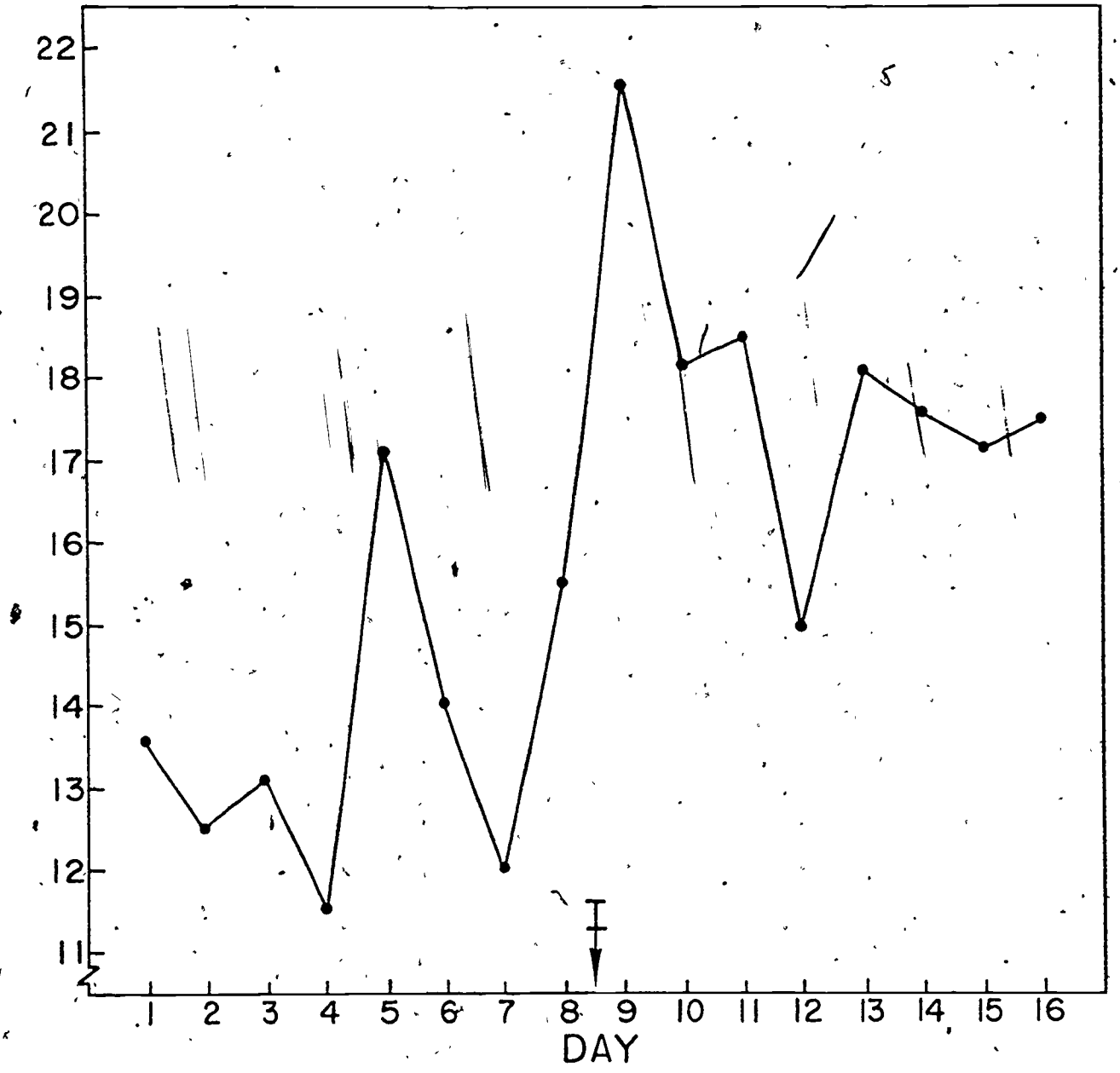


Figure 2. Mean cloze scores for a group of 45 fifth graders on 16 consecutive school days, with the last eight means artificially boosted by 5 points.

for detecting an intervention effect, will we be able to "retrieve" the built-in change in level of +5 units? If so, we may be reasonably assured that both the model chosen and the estimated parameter value must have been adequate.

The appropriate instance of the intervention-effect estimation technique developed by Glass and his coworkers, following Kepka (1972), is as follows. Using the random-shock form of the IMA(1,1) model equation (i.e., Eq. [35]) with k arbitrarily taken to be 0, we write

$$z_t = L_1 + (1-\theta_1) \sum_{i=1}^{t-1} a_i + a_t \quad (t=1,2,\ldots,8)$$

as the structural equation for observations from day 1 through day 8. (Here "observation" refers to the group mean for each day.) Then, after an intervention between days 8 and 9 which is assumed to result in a change of level by δ unit, the structural equation will change to

$$z_t + L_1 + (1-\theta_1) \sum_{t-1}^{t-1} a_i + a_t + \delta \quad (t=9,\ldots,16)$$

from day 9 on.

The next step is to recursively define a sequence of transformed variables $\{y_t\}$, as follows:

$$[42] \quad \begin{cases} y_1 = z_1 \\ y_t = (z_t - z_{t-1}) + \theta_1 y_{t-1}, \quad t \geq 2 \end{cases}$$

It can be shown that the y_t thus defined are expressible as linear functions of L_1, θ_1 and a_t. Namely,

$$y_1 = L_1 + a_1$$

$$y_2 = \theta_1 L_1 + a_2$$

$$\vdots$$

$$y_8 = \theta_1^7 L_1 + a_8$$

$$y_9 = \theta_1^8 L_1 + \delta + a_9$$

$$y_{10} = \theta_1^9 L_1 + \theta_1 \delta + a_{10}$$

$$\vdots$$

$$y_{16} = \theta_1^{15} L_1 + \theta_1^7 \delta + a_{16}$$

or, in matrix notation

$$[43] \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_8 \\ \hline y_9 \\ y_{10} \\ \vdots \\ y_{16} \end{bmatrix} = \left[\begin{array}{cc} 1 & 0 \\ \theta_1 & 0 \\ \vdots & \vdots \\ \theta_1^7 & 0 \\ \hline \theta_1^8 & 1 \\ \theta_1^9 & \theta_1 \\ \vdots & \vdots \\ \theta_1^{15} & \theta_1^7 \end{array}\right] \begin{bmatrix} L_1 \\ \delta \end{bmatrix} + \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_8 \\ \hline a_9 \\ a_{10} \\ \vdots \\ a_{16} \end{bmatrix}$$

which may symbolically be written

$$\underset{\sim}{y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{a},$$

where $\underset{\sim}{y}$ and $\underset{\sim}{a}$ are obvious, $\underset{\sim}{X}$ is the 16 x 2 matrix of successive powers of θ_1 and 0's, and $\underset{\sim}{\beta} = [L_1, \delta]'$.

Once the equation is cast in this form the standard least-squares estimate $\hat{\underset{\sim}{\beta}}$ of $\underset{\sim}{\beta}$ for linear models may be computed as

$$[44] \qquad \hat{\underset{\sim}{\beta}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}(\underset{\sim}{X}'\underset{\sim}{y}).$$

Here the vector $\underset{\sim}{y}$ is constructed, in accordance with Eqs. [42], from the "observed" sequence $\{z_t\}$ (which are the group means plotted in Figure 2), and the estimated $\hat{\theta}_1 = .5966$ replacing θ_1. We illustrate the calculations in detail for the first few elements of $\underset{\sim}{y}$. The observed $\underset{\sim}{z}$ vector is:

$$\underset{\sim}{z} = [13.56, 12.47, 13.11, 11.60, \ldots, 17.60, 17.22, 17.47]'$$

Hence, the vector of first differences is

$$\underset{\sim}{w} = [13.56, -1.09, .64, -1.51, \ldots, .53, -.38, .25]'$$

Then, in accordance with Eq. [42], we get

$$y_1 = z_1 = 13.56$$

$$y_2 = (z_2 - z_1) + \hat{\theta}_1 y_1$$
$$= -1.09 + (.5966)(13.56) = 6.9999$$

$$y_3 = .64 + (.5966)(6.9999) = 4.8161$$

$$y_4 = -1.51 + (.5966)(4.8161) = 1.3633,$$

and so on. The complete vector $\underset{\sim}{y}$ is, with elements rounded to two decimal places,

$$\underset{\sim}{y} = [13.56,\ 7.00,\ 4.82,\ 1.36,\ 6.28,\ .82,\ -1.64,\ 2.53$$
$$7.69,\ 1.21,\ .88,\ -2.94,\ 1.37,\ .29,\ -.21,\ .12]$$

With this and the 16 x 2 matrix $\underset{\sim}{X}$ with θ_1 replaced by its estimate $\hat{\theta}_1 = .5966$, we may compute $\hat{\beta}$ in accordance with Eq. [44]. The result is

$$\underset{\sim}{\hat{\beta}} = \begin{bmatrix} \hat{L}_1 \\ \hat{\delta} \end{bmatrix} = \begin{bmatrix} 13.2652 \\ 5.1276 \end{bmatrix}.$$

The estimated value, 5.1276, of δ is seen to be very close to the true value, 5.0, that we deliberately introduced into the system. Thus, we have some evidence to support the proposition that the model chosen and the estimated parameter value are adequate. This, in turn, suggests that the proposed method for estimating ρ_j is a viable one.

However, the skeptic may feel, in view of the artificial manner in which an "intervention effect" was introduced, that we merely "got out what we put in," and the particular value of $\hat{\theta}_1$ was immaterial. To check if this could have been the case, computations for $\hat{\delta}$ were repeated with the values of $\hat{\theta}_1$ used in Eqs. [42]-[44] systematically varied from .10 through .90 in steps of .05. The results, abbreviated to show the values of $\hat{\delta}$ only for every other $\hat{\theta}_1$ value used, were as follows:

$\hat{\theta}_1$	.10	.20	.30	.40	.50	.60	.70	.80	.90
$\hat{\delta}$	6.168	6.089	5.934	5.706	5.425	5.116	4.812	4.550	4.372

These results effectively refute the hypothetical skeptic's contention. The value used for $\hat{\theta}_1$ does make a difference in the value obtained for $\hat{\delta}$. And the value .5966 estimated by the proposed method comes close to being a optimal one. (By interpolation in the finer table, with $\hat{\theta}_1$ varied in steps of .05, the "best" value of $\hat{\theta}_1$ is found to be .6037, yielding $\hat{\delta} = 5.000$ to three decimal places.)

At the same time, however, we note that the obtained value of $\hat{\delta}$ varies fairly slowly with $\hat{\theta}_1$. In other words, the estimation of δ seems to be fairly robust with respect to minor inaccuracies in the estimation of θ_1. This is the ground on which we earlier asserted that further refinement of parameter estimates by maximum-likelihood methods seemed unnecessary, at least when the main purpose is to estimate the intervention effect. Of course, one instance does not prove a general proposition, and this assertion must remain a working hypothesis unless and until it is confirmed by further research.

Second Example: Simulated Data

In order to check the performance of the proposed method for a model of order higher than 1, simulated data following an AR(2) process were generated as follows.

Taking $\phi_1 = .6$, $\phi_2 = .3$ and $\beta = 3$ in Eq. [21], the particular AR(2) model used was

$$z_t = 3 + .6z_{t-1} + .3z_{t-2} + a_t,$$

with a_t generated by a random unit-normal generator and rescaled so that $\sigma_a = 4$. One hundred independent sequences

$$z_{i1}, z_{i2}, z_{i3}, \cdots, z_{i116}$$

were generated by use of the above equation, except for t = 1 and 2, for which

$$z_1 = 3 + a_1$$

and $$z_2 = 3 + .6z_1 + a_2$$

were used since there are no observations prior to z_1.

The result was a 100 x 16 data matrix $\underset{\sim}{Z}$, whose column means were as follows:

$\bar{z}_{.1} - \bar{z}_{.8}$: 18.71 22.41 22.14 22.96 23.94 24.79 24.90 24.94

$\bar{z}_{.9} - \bar{z}_{.16}$: 24.98 26.24 25.86 26.74 27.33 27.81 28.28 28.51 .

That these show a monotone increase with t reflects the fact our choice of σ_a (=4) was, in retrospect, too small relative to $\phi_1 = .6$, $\phi_2 = .3$ to produce an oscillitating series in the short run of 16 time points. This does not, however, vitiate the results of further analysis.

The correlation matrix based on this simulated data matrix is shown in Table 4, along with estimated autocorrelations of lags 1-15, calculated in accordance with Eq. [41].

Now let us pretend we did not know that these estimated autocorrelations were based on simulated data following a particular process, and go through the motions of identifying an appropriate model and estimating the parameter(s). First of all, we observe that there is no abrupt drop of the sample autocorrelations to near-zero; so an MA process is ruled out. Next, we note that there is a steady and fairly rapid declining of $\bar{r}_j$ with j--unlike the very gradual and irregular declining found in Table 2. So a stationary AR process of some order

Table 4. Correlation matrix based on a 100 x 16 simulated data matrix, each row of which follows the same AR(2) process

1.00															
.70	1.00														
.61	.66	1.00													
.56	.61	.73	1.00												
.46	.49	.68	.79	1.00											
.38	.43	.61	.72	.79	1.00										
.43	.41	.59	.63	.72	.79	1.00									
.42	.42	.54	.64	.69	.76	.85	1.00								
.32	.39	.46	.54	.60	.67	.80	.83	1.00							
.27	.30	.40	.45	.52	.58	.72	.80	.87	1.00						
.22	.27	.36	.39	.48	.54	.68	.75	.82	.87	1.00					
.20	.25	.30	.34	.39	.46	.62	.70	.75	.82	.85	1.00				
.11	.18	.26	.27	.36	.41	.55	.63	.68	.76	.83	.87	1.00			
.12	.15	.28	.25	.34	.39	.53	.58	.61	.69	.75	.81	.86	1.00		
.08	.16	.26	.24	.35	.37	.54	.57	.59	.67	.74	.74	.84	.88	1.00	
.06	.17	.22	.21	.28	.30	.50	.52	.48	.60	.69	.72	.78	.84	.89	1.00

AVERAGE LAGGED CORRELATIONS

Lag	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\bar{r}_j$	.816	.760	.685	.625	.560	.502	.442	.391	.332	.272	.239	.180	.166	.126	.057

is suggested (cf. Table 1 for the behavior of autocorrelations for various processes). The question is, what order?

The rate of decline does not seem quite as rapid as to suggest AR(1), which shows an exponential decay of the ρ_j. However, taking the successive ratios $\bar{r}_j / \bar{r}_{j-1}$ (which should all estimate ϕ_1 if an AR(1) model is adopted), it seems barely possible that an AR(1) model with $\phi_1 \approx .90$ might fit the data. (We say "barely possible" because the value .90 for ϕ_1 estimated from the successive ratios is considerably larger than $\bar{r}_1 = .816$, which should also be an estimate of ϕ_1 if AR(1) is in fact the correct model.) We therefore need to look at the estimated partial autocorrelations to decide the issue.

Setting $p = 2$, the Yule-Walker equations (cf. Eq. [25]) with ρ_1 and ρ_2 replaced by $\bar{r}_1$ and $\bar{r}_2$, respectively, are

$$\hat{\phi}_1 + .816\,\hat{\phi}_2 = .816$$

$$.816\,\hat{\phi}_1 + \hat{\phi}_2 = .760$$

whose solutions are

$$\hat{\phi}_1 = .586 \text{ and } \hat{\phi}_2 = .282.$$

Clearly, $\hat{\phi}_2$ is not small enough to conclude that $\phi_2 = 0$. That is, an AR(1) model is ruled out as inappropriate.

Next, let us compute $\hat{\phi}_{33}$ $(=\hat{\phi}_3)$ from the Yule-Walker equations with $p = 3$; i.e.,

$$\begin{bmatrix} 1 & .816 & .760 \\ .816 & 1 & .816 \\ .760 & .816 & 1 \end{bmatrix} \begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \hat{\phi}_3 \end{bmatrix} = \begin{bmatrix} .816 \\ .760 \\ .685 \end{bmatrix}$$

Using Cramer's rule, we get

$$\hat{\phi}_{33} = \frac{\begin{vmatrix} 1 & .816 & .816 \\ .816 & 1 & .760 \\ .760 & .816 & .685 \end{vmatrix}}{\begin{vmatrix} 1 & .816 & .760 \\ .816 & 1 & .816 \\ .760 & .816 & 1 \end{vmatrix}} = \frac{.0033}{.1028} = .032,$$

which is negligibly different from 0. We may therefore conclude that AR(2) offers an adequate fit to the data.

Once we adopt AR(2), our estimates of ϕ_1 and ϕ_2 are as previously computed from the Yule-Walker equations with p = 2; namely,

$$\hat{\phi}_1 = .586 \text{ and } \hat{\phi}_2 = .282.$$

Abandoning our make-belief that we do not know the "genealogy" of our data the estimated values for ϕ_1 and ϕ_2 are quite close to the actual values, .60 and .30, that were used to generate the simulated data. We may therefore conclude that the proposed method for parameter estimation "works" for second-order processes as well as the first.

SUMMARY AND REMARKS

The bulk of this chapter is admittedly expository in nature, but it is believed that the exposition was made in a more elementary manner than found in currently available books on the subject--although, by the same token, the treatment was necessarily incomplete in some technical detail.

The one original contribution made in this chapter was the

proposal of an alternative method for estimating autocorrelations of various lags--the key to model identification and parameter estimation in time-series analysis. This method is based on the ordinary sample correlation matrix which is computable whenever genuine longitudinal data are to be analyzed (i.e., when a single intact group has been observed at several time points). The traditional method for estimating autocorrelations (based on a single observation at each point in time, such as group means on the several measurement occasions), it was argued, is not appropriate for two reasons. First, it ignores the correlatedness inherent in longitudinal data, just as though we were to use a randomized groups design ANOVA when a repeated--measures design is proper. Second, the traditional method requires such a long series of observations in time (at least 50 observations, according to Box and Jenkins, 1970) as is almost never available in longitudinal studies.

The proposed method was put to a test by means of two numerical examples, one based on real data and the other, on simulated data. The outcomes of these analyses aquately confirmed the "validity" of the proposed method.

Directions for Future Research

Obviously, further study of the efficacy of the proposed method is needed; what was accomplished within the contract period has only scratched the surface in this respect. One thing which urgently needs to be done is to relax the assumption, inherent in the method as it stands, that the parameters are identical for all individuals in a group. This clearly an unrealistic assumption--although, in one sense, an innocuous one. When this assumption is untenable, what we

get as parameter estimates are some sort of averages of the respective individual parameters. However, it would be much more satisfactory if individual differences in the parameters could be explicitly considered. For instance, by assuming some particular distribution of each parameter over a population of individuals, the autocorrelations could probably be related to the moments of this distribution.

Another matter which requires further research is the method of estimating and testing intervention effects. The techniques developed by Glass, Willson and Gottman (1975) are perfectly satisfactory in situations where there is but one observation per time point. But, somehow, one feels that they are wasteful of information when applied to data from genuine longitudinal studies.

It is regrettable that the present researcher could make no inroads into the above-mentioned problems within the contrast period, mainly because he was a relative novice in the discipline of time-series analysis at the outset of the period--a novice who was dissatisfied with certain aspects of the traditional methods of time-series analysis when they are sought to be applied to longitudinal data. However, he intends to follow up this line of research in the future.

REFERENCES

Anderson, T. W. Estimation of covariance matrices which are linear combinations, or whose inverses are linear combinations, of given matrices. In Bose, R. C. et al. (eds.) Essays in probability and statistics. Chapel Hill: University of North Carolina Press, 1970.

Box, G.E.P. and Jenkins, G. M. Time-series analysis: Forecasting and control. San Francisco: Holden-Day, 1970.

Box, G.E.P. and Tiao, G. C. A change in level of a non-stationary time series. Biometrika, 1965, 52, 181-192.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Glass, G. V, Willson, V. L. and Gottman, J. M. Design and analysis of time-series experiments. Boulder, Colo.: Colorado Associated University Press, 1975.

Kepka, E. J. Model representation and the threat of instability in the interrupted time series quasi-experiment. Unpublished Ph.D. dissertation, Northwestern University, June 1972.

Nelson, C. R. Applied time series analysis for managerial forecasting. San Francisco: Holden-Day, 1973.

Quenouille, M. H. Approximate tests of correlation in time series. Journal of the Royal Statistical Society (Series B), 1949, 11, 68-84.

CHAPTER 9

ESTIMATION OF TRUE CHANGE: UPPER AND LOWER BOUNDS

INTRODUCTION

In Chapter 4 of this Report, Linn and Slinde have presented a survey of the literature on the topic of measurement of change and its many problems--seemingly insurmountable problems that led Cronbach and Furby (1970) to recommend against the use of gain scores, and advise instead that researchers "frame their questions in other ways."

Without discounting the seriousness of the problems surrounding the measurement of change, the present writers wish to propose that at least some of these problems can be traced to an unjustifiable assumption in classical test theory: that the error components of any pair of test scores are uncorrelated. In this chapter we explore new vistas that may be opened if the assumption of "universally uncorrelated measurement errors" is dropped. The dropping of this assumption, however, leads to mathematical problems that are insurmountable unless techniques hitherto not utilized in test theory--in particular, operator analysis--are introduced. This approach, pioneered in the first author's recent doctoral dissertation (K. Tatsuoka, 1975), is used in this chapter.

NOTATION AND DEFINITIONS

By and large, the notation used in this chapter follows that of Lord and Novick (1968), but there are some peculiarities. So we set forth a complete notational guide in this section, even though many of the symbols are in universal use and need no explanation.

All lower-case Roman letters (except those used as subscripts and superscripts) stand for person-space vectors in deviation form, rescaled by the factor $1/\sqrt{N-1}$, where N is the sample size. Thus, e.g.,

$$x = \left(\frac{X_1 - \bar{X}}{\sqrt{N-1}}, \frac{X_2 - \bar{X}}{\sqrt{N-1}}, \ldots, \frac{X_N - \bar{X}}{\sqrt{N-1}} \right)$$

is the N-vector whose elements are the deviation scores on test X for a sample of N persons, each divided by $\sqrt{N-1}$.

All Greek letters stand for scalars, while upper-case Roman letters either stand for scalars (like N) or are generic symbols for tests (like X and Y) or other random variables.

An immediate consequence of the above definition of the test vector x is that its squared norm (i.e., the scalar product of x with itself) represents the variance of test X:

$$(x,x) = \|x\|^2 = \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{\sqrt{N-1}} \right)^2 = \frac{\sum (X_i - \bar{X})^2}{N-1} = \sigma_x^2.$$

Similarly, the scalar product between two different test vectors x and y represents the covariance between tests X and Y:

$$(x,y) = \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{\sqrt{N-1}} \right) \left(\frac{Y_i - \bar{Y}}{\sqrt{N-1}} \right) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N-1} = \sigma(x,y).$$

Note that (x,y) is used instead of the more customary $x'y$ for a scalar product. This is because we will never have occasion to used the matrix product xy' of two vectors, and scalar products will mostly occur as coefficients in a linear combination of vectors so it is convenient to set them apart with parentheses.

In this notation the simple regression coefficient b_{yx} of Y on X, whose usual formula is

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{Cov(X,Y)}{Var(X)},$$

becomes

$$b_{yx} = \frac{\sigma(x,y)}{\|x\|^2} \text{ or simply } \frac{(x,y)}{\|x\|^2}$$

which further reduces to

$$b_{yx} = \sigma(x,y) \text{ or simply } (x,y)$$

when x is of unit norm (i.e., $\|x\| = 1$). This form will repeatedly occur in the sequel. Also, the correlation coefficient

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{\sum y^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

becomes

$$r_{xy} = \frac{\sigma(x,y)}{\|x\| \ \|y\|} = \frac{(x,y)}{\|x\| \ \|y\|}.$$

Hence, orthogonality of two vectors x and y [i.e., $(x,y) = 0$] is synonymous with the uncorrelatedness of the two tests X and Y which they represent ($r_{xy} = 0$). We shall often use the terms "orthogonal" and "uncorrelated" interchangeably--even though, strictly speaking, the former is a geometric property of two vectors while the latter is a statistical property of the two tests represented by the vectors.

The component of a vector y in the direction of another vector x is given by

$$(y,x)/\|x\|, \text{ or simply } (y,x) \text{ if } \|x\| = 1.$$

[This follows from the cosine law,

$$(x,y) = \|x\| \ \|y\| \cos \theta,$$

(where θ is the angle between the vectors x and y) and the fact, verifiable by elementary geometry, that the component in question is $\|y\| \cos \theta$.]

The projection (more precisely orthogonal) of a vector y onto vector x is a vector whose norm (length) is equal to the component of y in the direction x, and whose direction is that of x. In other words, it is the component (as defined above) multiplied by the unit vector in the direction of x; i.e.,

$$\text{Proj } (y|x) = \frac{(y,x)}{\|x\|} \cdot \frac{x}{\|x\|} = \frac{(y,x)}{\|x\|^2} x.$$

Note that the coefficient of x here is precisely the regression coefficient b_{yx} of y on x, defined earlier. Thus, the projection of y on x is the same thing as the regression of test Y on test X, and may be denoted

$$\hat{y} = R(y|x) = \frac{(y,x)}{\|x\|^2} x.$$

This interpretation of regression as the outcome of applying the "projection operator" to a vector is what enables us to utilize the various theorems and techniques of operator analysis alluded to in the Introduction.

The multiple regression of test Y on tests $X_1, X_2, \ldots, X_p$ is denoted by

$$\hat{y} = R(y|x_1, x_2, \ldots, x_p).$$

Geometrically, $\hat{y}$ corresponds to the projection of y onto the space spanned by $x_1, x_2, \ldots, x_p$.

Finally, two symbols which probably need no explanation are:

ρ_i = reliability of test X_i

and

$\rho(x,y)$ = correlation between X and Y.

ESTIMATING TRUE CHANGE FROM PRE- AND POST-TEST SCORES

The multiple regression equation for estimating $T_2 - T_1$ from the observed pre- and post-test scores, X_1 and X_2, may be written as

[1] $$t_2 - t_1 = R(t_2 - t_1 | x_1, x_2).$$

However, it is more convenient to use as predictors a pair of uncorrelated variables (such as the principal components, for example) instead the original X_1 and X_2 themselves. A further convenience is to have the derived predictor variables standardized so their vectors will be of unit norm. It is well-known that multiple regression is invariant of any nonsingular linear transformation of the predictor variables; i.e., if the derived predictors are linear combinations of the original predictors such that the coefficient determinant is non-zero, then using the multiple regression equation with the transformed predictors will yield predictions identical to those using the original multiple regression equation. For example, if the original predictors are X_1 and X_2, a new pair of predictors Y_1 and Y_2 defined by

$$Y_1 = \gamma_{11}X_1 + \gamma_{12}X_2$$

$$Y_2 = \gamma_{21}X_1 + \gamma_{22}X_2,$$

will leave the predictions unchanged so long as

$$\begin{vmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{vmatrix} \neq 0.$$

For the above reasons, we propose to replace Eq. [1] by an equivalent multiple regression equation using a pair of uncorrelated, unit-norm vectors $\{c_1, c_2\}$ (mathematically known as an orthonormal base of the space spanned by x_1 and x_2) as the predictors, i.e.,

$$[2] \qquad \widehat{t_2 - t_1} = R(t_2 - t_1 | c_1, c_2),$$

where the exact nature of c_1 and c_2 (i.e., how they are derived from x_1 and x_2) is to be specified later. Since c_1 and c_2 are uncorrelated and have unit norms (i.e., the standard deviations of C_1 and C_2 are unity), Eq. [2] may further be rewritten, successively, as

$$[3] \qquad \widehat{t_2 - t_1} = \sigma(t_2 - t_1, c_1)c_1 + \sigma(t_2 - t_1, c_2)c_2$$

$$= [\sigma(t_2, c_1) - \sigma(t_1, c_1)]c_1 + [\sigma(t_2, c_2) - \sigma(t_1, c_2)]c_2$$

[The first step follows from the facts that, when the predictors are uncorrelated, the partial regression coefficients are the same as the simple regression coefficients, and that c_1 and c_2 are of unit norm--see Section 2. The second step follows from the fact that the covariance of the difference between two variables with a third equals the difference between their respective covariances with the third variable: $\mathrm{Cov}(A-B,C) = \mathrm{Cov}(A,C) - \mathrm{Cov}(B,C)$.]

From the last member of Eq. [3] it is apparent that, in order to be able to use Eq. [2] in practice, we must know (i.e., be able to calculate)

$$\sigma(t_1,c_1),\ \sigma(t_1,c_2),\ \sigma(t_2,c_1) \text{ and } \sigma(t_2,c_2).$$

Recalling that c_1 and c_2 are to be defined as linear combinations of x_1 and x_2, i.e.,

$$c_i = \alpha_{i1} x_1 + \alpha_{i2} x_2 \quad (i=1,2),$$

it follows that

$$\begin{aligned}\sigma(t_j,c_i) &= \sigma(t_j,\alpha_{i1}x_1+\alpha_{i2}x_2)\\ &= \sigma(t_j,\alpha_{i1}x_1) + \sigma(t_j,\alpha_{i2}x_2)\\ &= \alpha_{i1}\sigma(t_j,x_1) + \alpha_{i2}\sigma(t_j,x_2)\ (i=1,2;\ j=1,2).\end{aligned}$$

Therefore, to use Eq. [2] we must know

$$\sigma(t_1,x_1),\ \sigma(t_1,x_2),\ \sigma(t_2,x_1) \text{ and } \sigma(t_2,x_2).$$

Of these, however, we already know the like-subscripted covariances, $\sigma(t_1,x_1)$ and $\sigma(t_2,x_2)$; i.e.,

[4] $$\sigma(t_1,x_1) = \|x_1\|^2\rho_1 \text{ and } \sigma(t_2,x_2) = \|x_2\|^2\rho_2,$$

where ρ_1 and ρ_2 are the reliabilities of the pretest X_1 and posttest X_2, respectively.[1]

---

[1] Each of Eqs. [4] may be derived as follows:

$$\rho_x = \rho(x,t)^2 = \left[\frac{\sigma(x,t)}{\sigma_x\sigma_t}\right]^2$$

$$\therefore\quad \sigma(x,t) = \sigma_x\sigma_t\sqrt{\rho_x}.$$

But $$\sqrt{\rho_x} = \frac{\sigma_t}{\sigma_x},\ \text{so } \sigma_t = \sigma_x\sqrt{\rho_x}$$

$$\therefore\quad \sigma(x,t) = \sigma_x(\sigma_x\sqrt{\rho_x})\sqrt{\rho_x} = \sigma_x^2\rho_x.$$

Hence, we need only show how to find the corss-subscripted covariances, $\sigma(t_1,x_2)$ and $\sigma(t_2,x_1)$.

It turns out that these connot be determined exactly, but their upper and lower bounds can be computed. Toward this end, we first discuss some mathematical preliminaries.

BOUNDS FOR (t_j,x_i) WHERE $(i \neq j)$

A powerful mathematical tool for obtaining bounds on scalar products of the sort we are interested in is Bessel's Inequality:

> Given an orthonormal set $\{a_1, a_2, \ldots, a_\nu\}$ (i.e., a set of mutually orthogonal vectors all of unit norm) and any vector y, it is true that

$$[5] \qquad \sum_{i=1}^{\nu} (y,a_i)^2 \leq \|y\|^2 .$$

It may be noted that, in any <u>finite</u> dimensional space, this inequality follows readily from the Pythagorean theorem. The equal sign holds when ν is the dimensionality of the space in which y lies (i.e., when $\{a_1, a_2, \ldots, a_\nu\}$ is a <u>complete</u> orthonormal set, or an orthonormal <u>base</u> of the space), for the sum on the left is then the sum of the squares of the components of y along all of the orthogonal axes. If ν is less than the dimensionality of the space, the left-hand sum will lack the squares of some of the components of y, and hence the "less than" sign may hold. (We cannot say that the "less than" sign <u>necessarily</u> holds, because the components whose squares are missing may happen to be zero anyway.) The reason why inequality [5] is given a celebrated name is that Bessel proved it to hold even for a vector space of infinite dimensionality (i.e., a Hilbert space), in which case ν itself may be

infinite and yet $\{a_1, a_2, \ldots\}$ may fail to be a complete orthonormal set.

For our particular application, we choose the orthonormal set $\{a_1, a_2, \ldots, a_\nu\}$ as follows: Let $x_1', x_1'', \ldots, x_1^{(N-2)}$ be the observed-score vectors of N - 2 parallel tests of X_1, and $e_1', e_1'', \ldots, e_1^{(N-2)}$ be the corresponding error-score vectors. Then, since the error components of any two parallel tests are by definition uncorrelated, it follows that

$$\{e_1^{(0)}/\|e_1^{(0)}\|, e_1'/\|e_1'\|, e_1''/\|e_1''\|, \ldots, e_1^{(N-2)}/\|e_1^{(N-2)}\|\}$$

is an orthonormal set comprising N - 1 vectors (one less than the total dimensionality, N, of our space). Here $e_1^{(0)}$ is the error-score vector of X_1 itself, the superscript '(0)' being added for consistency of notation.

Using this particular orthonormal set as the $\{a_1, a_2, \ldots, a_\nu\}$ in Bessel's inequality [5], we get

$$[6] \qquad \sum_{i=0}^{N-2} (y, e_1^{(i)}/\|e_1^{(i)}\|)^2 \leq \|y\|^2.$$

Now, from the definition of reliability, we know that

$$\|e_1^{(i)}\|^2 = \|x_1\|^2(1-\rho_1)$$

for all i = 0, 1, 2, ..., N - 2. Therefore [6] becomes

$$\sum_{i=0}^{N-2} (y, e_1^{(i)}/\|x_1\|\sqrt{1-\rho})^2 \leq \|y\|^2,$$

or, upon factoring out $1/\|x_1\|^2(1-\rho_1)$ from the summation on the left and dividing by it on both sides,

$$[7] \qquad \sum_{i=0}^{N-2} (y, e_1^{(i)})^2 \leq \|y\|^2 \|x_1\|^2(1-\rho_1).$$

This relation, as it stands, is clearly intractable. We therefore introduce a simplifying assumption: that the error component of each of several parallel tests has the same covariance with the error component of a given external test, or the assumption of "homogeneity of error coveriances" for parallel measures with another test, for brevity. Symbolically, we assume

[8] $$\sigma(e_y, e_1^{(0)}) = \sigma(e_y, e_1^{(1)}) = \ldots = \sigma(e_y, e_1^{(N-2)}) \equiv \sigma(e_y, e_1), \text{ say.}$$

This assumption is not as far-fetched as it may seem at first glance, for it merely requires that the observed-score covariances between Y and each of X_1, X_1', ..., $X_1^{(N-2)}$ are all equal.[2] Furthermore, $\sigma(y,x_1) = \sigma(y,x_1') = \ldots$, together with the assumption that $\sigma_{x_1} = \sigma_{x_1'} = \ldots$ (since X_1, X_1', ... are parallel measures), implies and is implied by

---

[2] This may be seen as follows:

$$\sigma(y,x_1) = \sigma(y,x_1')$$

$$\longrightarrow \quad \sigma(t_y+e_y, t_1+e_1) = \sigma(t_y+e_y, t_1+e_1')$$

[because any observed score is, by definition, equal to the sum of the true score and the error score, and since x_1 and x_1' have the same true-score component]

$$\longrightarrow \quad \sigma(t_y,t_1) + \sigma(t_y,e_1) + \sigma(e_y,t_1) + \sigma(e_y,e_1)$$
$$= \sigma(t_y,t_1) + \sigma(t_y,e_1') + \sigma(e_y,t_1) + \sigma(e_y,e_1')$$

$$\longrightarrow \quad \sigma(e_y,e_1) = \sigma(e_y,e_1')$$

[since $\sigma(t_y,e_1) = \sigma(t_y,e_1') = 0$]

$\rho(y,x_1) = \rho(y,x_1') = \ldots$. Thus, the homogeneity of error covariances assumption [8] is seen to be equivalent to assuming that all members of a set of parrallel tests correlate equally with a given external test, which seems to be a reasonable assumption.

It should be noted that [8] represents a liberalization of the traditional assumption in classical test theory, in that [8] merely states that the N - 1 error covariances are <u>equal</u> while the traditional assumption requires that these covariances all be <u>equal to zero</u> (the "universally uncorrelated measurement errors" assumption). In other words, the traditional assumption is a special case of [8], with $\sigma(e_y,e_1) = 0$.

When we introduce Eqs. [8] into inequality [7], the summands on the left all become equal, and the sum reduces to (N-1) (e_y,e_1). Hence, inequality [7] reduces to

$$[9] \qquad (y,e_1)^2 \leq \|y\|^2 \, \|x_1\|^2 \, \frac{1 - \rho_1}{N - 1}.$$

Note, incidentally, that this implies that if $\rho_1 = 1$ or $N \to \infty$, $(y,e_1) = 0$—in agreement with the traditional assumption. It is clear, however, that the "homogeneity of error covariances" assumption [8] is incompatible with letting $N \to \infty$, for then the infinite series on the left-hand side of inequality [7] must diverge (since it is the sum of an infinite number of constant positive terms) and cannot be bounded. We therefore exclude the possibility that $N \to \infty$, and conclude that the only condition under which [9] leads to the classical assumption, $(y,e_1) = 0$, is when $\rho_1 = 1$. That is, <u>within the realm of perfectly reliable tests</u>, the error components of any two tests are always

uncorrelated--which is trivially true since the error scores are constantly equal to zero anyway.

Next, from the definition

$$x_1 = t_1 + e_1$$

it follows that

$$e_1 = x_1 - t_1$$

and hence that

$$(y,e_1) = (y,x_1) - (y,t_1).$$

Substituting this in [9], we get

$$[(y,x_1) - (y,t_1)]^2 \leq \|y\|^2 \, \|x_1\|^2 \, \frac{1 - \rho_1}{N - 1},$$

or

$$-\|y\| \cdot \|x_1\| \sqrt{\frac{1-\rho_1}{N-1}} - (y,x_1) \leq (y,t_1) \leq \|y\| \cdot \|x_1\| \sqrt{\frac{1-\rho_1}{N-1}},$$

whence

[10] $$(y,x_1) - \|y\| \cdot \|x_1\| \sqrt{\frac{1-\rho_1}{N-1}} \leq (y,t_1) \leq (y,x_1) + \|y\| \cdot \|x_1\| \sqrt{\frac{1-\rho_1}{N-1}}.$$

Note, again that if $\rho_1 = 1$, this yields

$$(y,t_1) = (y,x_1),$$

which is the classical test-theory result under the assumption of uncorrelated errors of measurement for any pair of tests.

Now, recalling that y was an arbitrary test vector (other than one of the parallel measures of x_1), we may let $y = x_2$, the post-test

vector. In this instance [10] becomes

$$[11a] \quad (x_1,x_2) - \|x_1\| \cdot \|x_2\| \sqrt{\frac{1-\rho_1}{N-1}} \leq (t_1,x_2) \leq (x_1,x_2) + \|x_1\| \cdot \|x_2\| \sqrt{\frac{1-\rho_1}{N-1}}$$

and, similarly, by interchanging the roles of x_1 and x_2, we get

$$[11b] \quad (x_1,x_2) - \|x_1\| \cdot \|x_2\| \sqrt{\frac{1-\rho_2}{N-1}} \leq (t_2,x_i) \leq (x_1,x_2) + \|x_1\| \cdot \|x_2\| \sqrt{\frac{1-\rho_2}{N-1}}.$$

Thus, we have established upper and lower bounds for $\sigma(t_1,x_2)$ and $\sigma(t_2,x_1)$, the cross-subscripted covariances which were all that remained to be known in order to be able to use Eq. [2] in practice. It is true that we have not determined these covariances exactly (which seems impossible to do in principle), and hence an exact estimate of $t_2 - t_1$ is infeasible. However, by suitable substitutions of the upper and lower bounds of $\sigma(t_j,x_i)$--depending on whether they appear with a positive or negative sign in the regression equation after c_1 and c_2 have been specified--we are able to obtain upper and lower bounds for $\widehat{t_2 - t_1}$.

A computer program for implementing the foregoing developments is being written, but it could not be completed within the contract period--mainly because it seeks to permit a larger set of predictor variables than just $\{x_1,x_2\}$ in estimating $t_2 - t_1$. For it stands to reason (as, indeed, Cronbach and Furby, 1970, have suggested) that the more predictors--including demographic variables--we employ, the better will be the accuracy with which we can estimate $t_2 - t_1$.

As this point, we can only present computed results for a

lower bound of the accuracy of the estimate $\widehat{t_2 - t_1}$, to which we address ourselves in the next section.

ACCURACY OF ESTIMATE

The accuracy of any estimate made by multiple regression may be guaged by the multiple correlation coefficient. In the present context, we wish to calculate $\rho(\widehat{t_2-t_1},\ t_2-t_1)$, where $\widehat{t_2 - t_1}$ is defined by Eq. [2]. However, since its exact value cannot be determined in principle, we must be satisfied with finding a lower bound for $\rho(t_2-t_1,\ t_2-t_1)$.

It is well-known that, when the predictor variables are uncorrelated, the squared multiple -R is the sum of the squares of the zero-order correlations between the several predictors and the criterion. For the case at hand, we have

$$\rho^2(\widehat{t_2-t_1},\ t_2-t_1) = \rho^2(t_2-t_1,c_1) + \rho^2(t_2-t_1,c_2),$$

or, since c_1 and c_2 are of unit norm besides being orthogonal (uncorrelated),

$$[12] \qquad \rho^2(\widehat{t_2-t_1},\ t_2-t_1) = \frac{\sigma^2(t_2-t_1,c_1)}{\|t_2-t_1\|^2} + \frac{\sigma^2(t_2-t_1,c_2)}{\|t_2-t_1\|^2}$$

Here $\{c_1,c_2\}$ may be any orthonormal base of the space spanned by x_1 and x_2. It is natural to take as c_1 the unit vector in the direction $x_2 - x_1$ (since we are estimating $t_2 - t_1$), whereupon c_2 is the unit vector orthogonal to $x_2 - x_1$ in the plane defined by x_1 and x_2. This procedure for constructing an orthonormal base is called the Gram-Schmidt procedure (see, e.g., Rao, 1968). The results are

$$c_1 = (x_2 - x_1)/\|x_2 - x_1\|$$

$$c_2 = \{x_2 - (x_2, c_1)c_1\}/\|x_2 - (x_2, c_1)c_1\|.$$

With this special choice of c_1 and c_2 (recall that any non-singular linear transformation of x_1 and x_2 will leave the multiple regression, and hence also the multiple correlation coefficient, invariant), the two terms on the right-hand side of Eq. [12] acquire the following interpretations:

First term = reliability[3] of $X_2 - X_1$

Second term = squared correlation between $T_2 - T_1$ and the residualized post-test score, partialling out $X_2 - X_1$.

---

[3]Because, by definition,

$$\rho_{x_2-x_1} = \rho^2_{t_2-t_1,\ x_2-x_1}$$

$$= \frac{\sigma^2(t_2-t_1,\ x_2-x_1)}{\sigma^2_{t_2-t_1}\ \sigma^2_{x_2-x_1}}$$

$$= \frac{\sigma^2(t_2-t_1,\ x_2-x_1)}{\|t_2-t_1\| \cdot \|x_2-x_1\|^2}$$

$$= \frac{\sigma^2(t_2-t_1,\ (x_2-x_1)/\|x_2-x_1\|)}{\|t_2-t_1\|^2}$$

Since c_1 and c_2 are linear combinations of x_1 and x_2, the numerators of the fractions on the right-hand side of [12] are quadratic functions of $\sigma(t_1,x_1)$, $\sigma(t_1,x_2)$, $\sigma(t_2,x_1)$, $\sigma(t_2,x_2)$, of which the like-subscripted covariances are, as mentioned earlier, known exactly, and we have obtained upper and lower bounds for the cross-subscripted covariances as inequalities [11a] and [11b] above. Hence, lower bounds of these numerator expressions may be calculated by substituting the lower or upper bounds of (t_1,x_2) and (t_2,x_1)--depending on the signs with which they occur.

The denominator expression (common to both fractions) does not immediately appear to be related to (t_1,x_2) and (t_2,x_1), but a little algebraic manipulation reveals that it actually is related to them. To wit,

[13] $$\|t_2-t_1\|^2 = (t_2-t_1,\ t_2-t_1)$$

$$= \|t_2\|^2 + \|t_1\|^2 - 2(t_1,t_2)$$

$$= \|x_2\|^2\rho_2 + \|x_1\|^2\rho_1 - 2(t_1,t_2),$$

the first two of the three terms of the last expression being directly observable. But

$$(t_1,t_2) = (t_1,x_2-e_1)$$

$$= (t_1,x_2), \text{ since } (t_1,e_1) = 0.$$

Similarly,

$$(t_1,t_2) = (x_1,t_2).$$

To get a lower bound for $\rho^2(\widehat{t_2-t_1},\ t_2-t_1)$, we need an upper bound of the denominator $\|t_2-t_1\|^2$, and hence a lower bound of (t_1,t_2), for this occurs with a negative sign in expression [13] for $\|t_2-t_1\|^2$. Since (t_1,t_2) is equivalently equal to (t_1,x_2) and to (x_1,t_2), as shown above (but not equal to (x_1,x_2) unless the "universally uncorrelated measurement errors" assumption is invoked), we must use min $\{\ell.b.(t_1,x_2),\ \ell.b.(x_1,t_2)\}$--i.e., the smaller of the lower bounds of (t_1,x_2) and (x_1,t_2)--to replace (t_1,t_2) in expression [13].

The foregoing completes our outline of how a lower bound of $\rho^2(\widehat{t_2-t_1},\ t_2-t_1)$ may be computed. Details of the computation are carried out by a computer program.[4] We now turn to a numerical example utilizing real data. This example not only illustrates the actual calculations for the above developments, but shows how we may introduce other predictors besides the pre- and post-tests themselves in order to increase the accuracy of estimating $\widehat{t_2 - t_1}$.

NUMERICAL EXAMPLE

The data for this example are from an unpublished study by Misselt (1973), in which (among other things) the Metropolitan Achievement Test battery was administered to a large group of third graders in the Champaign, Illinois school district in the school year 1971-72. The group was retested in 1972-73 as fourth graders. Only the Reading test in the battery is considered below, and only the scores for 624

---

[4]Available on request from the authors. This program accommodates three other variables besides the pre- and post-tests themselves.

pupils who took the test both in 1971-72 ("pretest") and in 1972-73 ("posttest") are utilized. Besides the pretest and posttest scores in reading, IQ scores were available for these pupils, so IQ was used as a third variable in the computations that follow.

We therefore extend Eq. [2] to

$$[14] \qquad \widehat{t_2-t_1} = R(t_2-t_1 | c_1, c_2, c_3)$$

$$= (t_2-t_1, c_1)c_1 + (t_2-t_1, c_2)c_2 + (t_2-t_1, c_3)c_3,$$

where c_1, c_2 and c_3 are constructed by the Gram-Schmidt procedure as

$$c_1 = (x_2-x_1)/\|x_2-x_1\|$$

$$c_2 = \{x_2-(x_2,c_1)c_1\}/\|x_2-(x_2,c_1)c_1\|$$

$$c_3 = \{x_3-(x_3,c_1)c_1 - (x_3,c_2)c_2\}/\|x-(x_3,c_1)c_1-(x_3,c_2)c_2\|$$

Eq. [12] for the squared multiple correlation, $\rho^2(\widehat{t_2-t_1}, t_2-t_1)$, is accordingly generalized to

$$[15] \qquad \rho^2(\widehat{t_2-t_1}, t_2-t_1) = \frac{\sigma^2(t_2-t_1,c_1)}{\|t_2-t_1\|^2} + \frac{\sigma^2(t_2-t_1,c_2)}{\|t_2-t_1\|^2} + \frac{\sigma^2(t_2-t_1c_3)}{\|t_2-t_1\|^2}$$

Summary statistics for the three tests and some intermediate results necessary for calculating $\rho^2(\widehat{t_2-t_1}, t_2-t_1)$ when the assumption $\rho(e_1,e_2) = 0$ is invoked, and its lower bound when this assumption is not used, are shown in Table 1.

Table 1. Intermediate results needed for calculating $\rho^2(\widehat{t_2-t_1}, t_2-t_1)$.

	Mean	s.d.	ρ_i	$\sqrt{\frac{1-\rho_i}{N-1}}$ (N = 624)
Reading Pretest (X_1)	27.82	10.92	.95	.00895
Reading Posttest (X_2)	35.12	12.41	.95	.00895
IQ (X_3)	104.24	18.75	--	--

Covariance matrix for X_1, X_2, X_3:

$$\begin{bmatrix} 119.19 & 113.02 & 137.76 \\ 113.02 & 153.90 & 163.67 \\ 137.76 & 163.67 & 351.51 \end{bmatrix}$$

The covariances (t_j, c_i); [j=1,2;i=1,2,3], under the assumption that $(e_1, e_2) = 0$:

$$\begin{bmatrix} .0304 & 10.3876 & .2391 \\ -4.8384 & 13.4378 & -2.2270 \end{bmatrix}$$

Normalizing divisors for c_1, c_2, c_3:

$$K_1 = \|x_2 - x_1\| = 6.8592$$

$$K_2 = \|x_2 - (x_2, c_1)c_1\| = 10.8801$$

$$K_3 = \|x_3 - (x_3, c_1)c_1 - (x_3, c_2)c_2\| = 12.9970$$

Based on the intermediate results displayed in Table 1, we first calculate the bounds for $\sigma(t_1, x_2)$ and $\sigma(t_2, x_1)$, and note that when the assumption $\sigma(e_1, e_2) = 0$ (an instance of the "universally uncorrelated measurement errors" of classical test theory) is invoked,

$$\sigma(t_1,x_2) = \sigma(t_2,x_1) = \sigma(x_1,x_2).$$

From inequality [11a] we get

$$113.02 - (10.92)(12.41)(.00895) \leq (t_1,x_2) \leq 113.02 + (10.92)(12.41)(.00895)$$

or

$$111.81 \leq (t_1,x_2) \leq 114.23$$

when the traditional assumption $\sigma(e_1,e_2) = 0$ is <u>not</u> invoked. Whereas

$$\sigma(t_1,x_2) = \sigma(x_1,x_2) = 113.02$$

When we assume $(e_1,e_2) = 0$.

In this numerical example, since $\rho_1 = \rho_2$ (=.95), the bounds for $\sigma(t_2,x_1)$ are exactly the same as those for $\sigma(t_1,x_2)$, as is evident by comparing inequalities [11a] and [11b]. This will not be true in general, when $\rho_1 \neq \rho_2$. Of course, under the classical assumption that $\sigma(e_1,e_2) = 0$, $\sigma(t_1,x_2)$ and $\sigma(t_2,x_1)$ are always the same, both being equal to $\sigma(x_1,x_2)$.

Before calculating the lower bound for $\rho(\widehat{t_2-t_1}, t_2-t_1)$ under the liberalized assumption of "homogeneity of error covariances" for parallel measures, let us calculate the exact value of $\rho(\widehat{t_2-t_1}, t_2-t_1)$ which the classical assumption of universally uncorrelated measurement errors purports to enable us to get. Note that, under this assumption, the common denominator of the fractions on the right-hand side of Eq. [15] can be exactly computed from Eq [13]:

$$\|t_2-t_1\|^2 = \|x_2\|^2\rho_2 + \|x_1\|^2\rho_1 - 2(t_1,t_2)$$

$$= (153.90)(.95) + (119.19)(.95) - (2)(113.02)$$

$$= 33.3955.$$

Then, using the intermediate results displayed in Table 1, we get the following values for the three terms on the right-hand side of Eq. [15], whose sum should equal $\rho^2(\widehat{t_2-t_1},\ t_2-t_1)$:

First term (reliability of $X_2 - X_1$	=	.7098
Second term	=	.2781
Third term	=	.1821
TOTAL		1.1705

This result is, of course, absurd since $\rho^2(\widehat{t_2-t_1},\ t_2-t_1)$ cannot exceed unity. This is but one instance of the various difficulties that arise from the traditional assumption of universally uncorrelated measurement errors. (See K. Tatsuoka, 1975, for other examples.)

We now turn to the calculation of a lower bound for $\rho(t_2-t_1,\ t_2-t_1)$ under the liberalized assumption of homogeneity of error covariances for parallel measures. Table 2 shows the intermediate results necessary for this purpose, in addition to or in lieu of the values displayed in Table 1.

Table 2. Intermediate results needed calculating a lower bound for $\rho(\widehat{t_2-t_1},\ t_2-t_1)$ in the absence of the assumption $\rho(e_1,e_2) = 0$

Lower and upper bounds for $\sigma(t_j,c_i)$:					
- .1466	10.1793	.1911	.2071	10.5960	.2871
-5.0151	13.3410	-3.0491	-4.6615	13.5347	-1.4050

Table 2 (Continued)

Lower and upper bounds for $\|t_2-t_1\|^2$, from Eq. [13] and the bounds for (t_1,t_2):
$30.9687 \leq \|t_2-t_1\|^2 \leq 35.8202$

Based on these intermediate results, we find the lower-bound values of the three terms on the right-hand side of Eq. [15] to be:

First term (reliability of $X_2 - X_1$)	$\geq$	.5691
Second term	$\geq$	.2136
Third term	$\geq$	.0711
$\therefore \quad \rho^2(\widehat{t_2-t_1},\ t_2-t_1)$	$\geq$	.8538

Hence, a lower bound of the multiple correlation $\rho(\widehat{t_2-t_1},\ t_2-t_1)$, a measure of the accuracy of estimating $t_2 - t_1$ by the method proposed in this chapter is,

$$\sqrt{.8538} = .9223.$$

SUMMARY

A vector-geometric and operator-analytic approach to derivations and proofs in test theory, first explored by K. Tatsuoka in her dissertation (1975), was applied in this chapter to the problem of estimating the true change from pre- to post-tests. One advantage of this approach is that it renders feasible hitherto intractable mathematical developments in the absence of the traditional simplifying assumption that error scores are universally uncorrelated.

That this assumption is inadmissible as an universal postulate

has been argued--with examples of "paradoxes" to which it leads--by K. Tatsuoka (1975). Linn and Slinde have also pointed out, in Chapter 4 of this Report, that--especially in the case when pre- and post-tests are under consideration--the assumption of uncorrelated errors is unjustifiable.

Upper and lower bounds for estimated true change were developed without the uncorrelated errors assumption, but with the less restrictive assumption that the error covariances of a set of parallel tests with an external variable are all equal (the "homogeneity of error covariances" assumption.) In addition, a lower bound for the multiple correlation $\rho(\widehat{t_2-t_1}, t_2-t_1)$ between estimated true change and actual change was derived. It was also noted that, under the traditional uncorrelated errors assumption, not only a lower bound, but the actual correlation value could be computed. When this was done for the numerical example (using real data), however, a value exceeding unity was found--thus providing another piece of evidence of the inadmissibility of the universally uncorrelated errors assumption. With the relaxed assumption, a reasonable and useful lower bound (.9223) was obtained.

REFERENCES

Cronbach, L. J., & Furby, L. How we should measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80. See also Errata, Ibid., 1970, 74, 218.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Misselt, L. A. An analysis at achievement level and achievement gains in the Champaign public schools. Unpublished paper, University of Illinois, 1973.

Rao, C. R. Linear statistical inference and its applications. New York: John Wiley, 1968.

Tatsuoka, K. K. Vector-geometric and Hilbert-space reformulations of classical test theory. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1975.

APPENDIX A

COMPARABLE READING TEST SCORES: A REVIEW OF THE ANCHOR TEST STUDY

Bianchini, J. C. & Loret, P. G. Anchor Test Study: Final Report. Report and Volumes 1 though 30, available as ERIC Documents ED 092 601 through ED 092 631.

Bianchini, J. C. & Loret, P. G. Anchor Test Study Supplement Final Report. Volumes 31 through 33, available as ERIC Documents. ED 092 632 through ED 092 634.

The prospect of reviewing the mammoth report of the Anchor Test Study (ATS) initially struck me as an overwhelming task. With the limited space in my office it would have been easy to refuse the request to review the ATS had it not been for the availability of microfiche. Although I haven't seen it in that form, hard copy of the 34 volumes of the final report requires about 8-1/2 feet of shelf space (Loret, 1974). An acquisition of that magnitude would require me to part with more of those dusty "should read sometime" items on my shelves than my conscience would allow. For better or worse, however, modern technology which made possible the production of the over 15,000 page report containing more than 8,000 computer produced tables and graphs in the first place also deprived me of my best alibi by reducing the report to a microfiche file that is only 2-3/4 inches thick.

Fortunately the task of reviewing the ATS for this journal was greatly simplified by the fact that a very good review of the ATS has already appeared in another NCME publication. The summer 1973 issue of Measurement in Education was devoted to a description of the study (Jaeger, 1973). Jaeger's description appeared more than a year before the full report was released and before the supplement study involving an eighth test was available. In addition to having directed the development of study specification, he had available at that time, all but the three volumes that comprise the supplement report. Indeed the 31-volume final report of the original study was delivered to USOE in December, 1972. The delay of almost two years between delivery of the report and its release is unfortunate because the value of norms certainly does not improve with age.

Jaeger's description of the ATS provides a good review of the history of the study, the planning and conduct of the study as well as the major outcomes of the study. A more recent overview of the study has been provided by the project director, Peter Loret (1974). Due to the availability of these two descriptions of the study I will try to keep my comments about the history and study procedures relatively brief.

OBJECTIVES AND BACKGROUND OF THE STUDY

"The Anchor Test Study had two major objectives: to provide a method by which one may translate a child's score on any one of seven widely used standardized reading tests into a score on any of the other tests, and to provide new nationally representative norms for each of these seven tests" (ATS. Final Report. Project Report, p. 1). This was subsequently expanded to eight tests but otherwise this concise statement of objectives needs no revision. Certainly there were other lesser objectives such as the empirical investigations of different equating techniques, and obtaining intercorrelations among the various tests, but these are minor in comparison to the two major objectives.

As noted by Jaeger (1973) and by Loret (1974) the concerns that led to the ATS have a long history. Dual concerns about the adequacy of national norms provided by test publishers and the desirability of being able to compare scores obtained on one test with those obtained on another have been with us for a long time (see for example Cureton, 1941; Lennon, 1964b).

The differences in sampling procedures that have been used by different publishers were clearly documented by Lennon (1964b). Even without differences in initial procedures, however, the relatively low rate of cooperation among selected schools that is enjoyed by publishers would make the representativeness of the norms questionable. The lack of representativeness and comparability creates difficulties when schools or school systems change from one battery to another or when an attempt is made to interpret scores of transfer students. Such difficulties, however, were not sufficient to motivate a major norming and equating study across several publishers.

There are many technical and political obstacles to equating tests across publishers (see Angoff, 1964; Flanagan, 1964; Lennon, 1964a; Lindquist, 1964). A strong motivation was needed to attempt to overcome these obstacles. This motivation was provided by the increasing demand for evaluations at the state and national level that occurred during the latter part of the 1960's. Early attempts to obtain achievement test data for the national evaluation of Title I, for example, were faced with a hodgepodge of different tests with different norms and different scales (Loret, 1974).

A major technical problem in equating tests of different publishers is that the tests may not measure the same characteristic. Angoff (1971), lists two requirements for equating, the first of which is that the "...instruments in question must measure the same characteristic..." (p. 573). With different content specifications used by different publishers, the satisfaction of this requirement seemed dubious. Intercorrelations among the tests obtained in a pilot study were found to be high enough, however, to make the equating seem worthwhile (Jaeger, 1973).

METHODOLOGY

The study was designed with two major phases: the norming phase and the equating phase. The norming was designed to provide national norms for individual pupils and for school means. The norms were developed for the vocabulary and the reading comprehension subtests as well as total reading for the Metropolitan Achievement Test, 1970 edition (MAT). The data were collected in April 1972 at grades 4, 5 and 6 and hence provide spring norms at those grade levels.

The sampling design for the norming study was developed by Westat Research, Inc. The design called for a stratified, random sample of 940 schools. The norms needed to be as representative of the nation's 4th, 5th and 6th-grade students as possible and great care and effort was devoted to the design of the sample. Primary-sample schools were selected, and for each school in the primary sample five schools with the same sampling characteristics were randomly selected as secondary sample schools, to use in place of non-participants in the primary sample. Due to careful planning and advance work with the Council of Chief State School Officers and others, relatively little reliance had to be placed on the secondary-sample schools (838 primary sample and 80 secondary sample schools with a total of approximately 65,000 pupils actually participating in the study). The high participation rate is a real tribute to the many people involved in the planning and conduct of the study. It also greatly enhances the value of the norms by minimizing the bias due to non-cooperation and is undoubtedly the single most important distinction of the study norms in comparison to the publishers' norms.

The equating phase of the study was designed to provide raw score equivalences for total reading, the vocabulary subtest and the reading comprehension subtest of seven major test batteries. Subsequently an eighth test was equated to the original seven in a study conducted in the spring of 1973. The tests, forms and the levels used at each grade level are summarized in Table 1. By equating of each of the other tests to the MAT (the anchor test) the norms obtained for the MAT were translated to norms for each of the other tests.

The sample characteristics for the equating phase are less crucial than in the norming phase of the study but again this phase of the study achieved a very high participation rate. Usable equating data were obtained in April 1972 for a total of almost 135,000 students for the original seven tests. To equate the CMT to the anchor test and through it to the other six tests, usable data were obtained for another 14,400 students in April 1973.

The design of the administration of tests in the equating phase called for a sample of students to take each pair of tests in order AB and a sample in order BA. A schematic representation of the equating design is shown in Table 2. As can be seen in Table 2, in addition to the pairing of each test with every other test in both

Table A-1

Tests, Forms and Levels Used in the Anchor Test Study

	Abbreviated Title	Form	Level Used For Grade 4	5	6
California Achievement Tests (1970 ed.)	CAT	A	3	3	4
Comprehensive Tests of Basic Skills (1968 ed.)	CTBS	Q	2	2	3
Gates-MacGinitie Reading Tests (1964 ed.)*	GMT	1M	Survey D	Survey D	Survey D
Iowa Tests of Basic Skills (1971 ed.)	ITBS	5	10	11	12
Metropolitan Achievement Tests (1970 ed.)	MAT	F	Elementary	Intermediate	Intermediate
Sequential Tests of Educational Progress STEP Series II (1969 ed.)	STEP	A	4	4	4
SRA Achievement Series (1971 ed.)	SRA	E	Blue	Blue	Green
Stanford Achievement Tests (1964 ed.)	SAT	W	Intermediate I	Intermediate II	Intermediate III

*Not one of the seven tests in the original 1972 data collection. The GMT was equated to the other seven tests in a separate study conducted in 1973.

Table A-2

Schematic Representation of Equating Study Design

Test Administration Order (April 1972)

Test	1	2	3	4	5	6	7
1. CAT	1-1* 1*-1	1-2	1-3	1-4	1-5	1-6	1-7
2. CTBS	2-1	2-2* 2*-2	2-3	2-4	2-5	2-6	2-7
3. ITBS	3-1	3-2	3-3* 3*-3	3-4	3-5	3-6	3-7
4. MAT	4-1	4-2	4-3	4-4* 4*-4	4-5	4-6	4-7
5. STEP	5-1	5-2	5-3	5-4	5-5* 5*-5	5-6	5-7
6. SRA	6-1	6-2	6-3	6-4	6-5	6-6* 6*-6	6-7
7. SAT	7-1	7-2	7-3	7-4	7-5	7-6	7-7* 7*-7

Test Administration Order (April 1973)

Test	8	4
8. GMT	8-8* 8*-8	8-4
4. MAT	4-8	

*Indicates an alternate form of the test

possible orders, each test was also paired with its own alternate form in both an AB and a BA order. This provided for parallel-form reliability estimates for each test.

Eight combinations of two equating methods (linear and equipercentile) and four equating procedures (involving the use of different subsets of the data from the design shown in Table 2) were used to equate each pair of tests. These combinations of method and procedures were compared to each other and also evaluated in terms of estimated errors of equating. Based on these results, the equipercentile method and a procedure that involves pooling all the data for a given test for each order of administration and then averaging the equating results were found to be most satisfactory.

Following the equating of raw scores on all of the tests the percentile norms for individual pupils and for school means were obtained from the MAT norming study results. Comparisons of these norms to the norms provided by the publishers were then provided. Finally, the adequacy of the equating for several subgroups of students was investigated.

THE REPORT

Despite the voluminous nature of the ATS report readers should have relatively little difficulty in obtaining desired information from it regardless of the level of detail that is required. The needs of most users are amply met in a 92-page separate report entitled "Anchor Test Study: Equivalence and Norms Tables for Selected Reading Tests" which is available from the U.S. Government Printing Office as stock number 1780-01312 at a cost of $1.90. This report contains a brief description of the study and the primary tables that resulted from the study. The tables are divided into four major categories: equivalency tables, tables of individual score norms, tables of school mean norms, and a table that presents a comparison ATS percentile ranks with the corresponding percentile ranks from the publishers' norms.

For the reader who desires more technical detail the two volumes containing the "project reports" will usually suffice. These volumes which have the catchy titles, "Anchor Tests Study. Final Report. Project Report" and "Anchor Test Study Supplement. Final Report. Volume 31, Project Report" may be obtained from ERIC as documents ED 092 601 and ED 092 632 respectively. These reports contain detailed descriptions of the study methodology including the sampling, estimation and equating procedures. They also contain a discussion of the major results and technical evaluations of the study results. At this level the reader may also want to skim through some of the tables and graphs in Volumes 2 through 27 as well as those in 30, 32 and 33 to evaluate the adequacy of the summary and description of results in the project reports. I think that a small sampling of

those tables and graphs will impress most readers with the thoroughness and scrupulous accuracy of reporting in the project reports.

For anyone who wants to dig beyond the project reports I can only say that the tables and graphs are available through ERIC in quantitites that should satisfy even the most heary of appetites. Volumes 2 through 4* provide equating tables for the 8 combinations of methods and procedures, in addition to estimated errors of equating, and test intercorrelations for grades 4, 5 and 6 respectively. Volumes 5* through 10 provide graphs which compare the equating lines for different procedures and for different equating methods at each grade. Volumes 11 through 21 present subgroup equating tables (boys, girls, 3 IQ groups, 3 racial groups, and 3 SES groups). Graphs comparing the subgroup equating results to each other and to those for the total group are presented in Volumes 22 through 27. Volume 30 presents a comparison of the ATS norms with those provided by the test publishers, and reports conditional errors of equating, (i.e., the standard deviation of observed scores on test j around the equivalent score of test j for each value of test j') quality control results and information on the convergence of equating iterations.

The information in the first 30 volumes and in the project report is all concerned with the 7 reading achievement tests that were in the original study. (See Table 1.) The Supplement Report (Volumes 31 through 33) gives results of a study conducted a year after the original study for the purpose of equating an eighth test (the Gates McGinitie) to the original seven.

SELECTED RESULTS

MAT Norms

The norms that were obtained for the reading test of the MAT are probably the best national norms that have ever been obtained for a standardized achievement test. As already noted the school cooperation rate was exceptional. The sample design and weighting procedures were of very high technical quality.

---

*Although it is unlikely to cause anyone any real difficulty, it might be noted that the tables that belong in Volume 4 have been inadvertently put on the Volume 5 microfiche (ED 092 606) under the title "Equating Procedure Comparison Graphs, Grade 4". The graphs that belong in Volume 5 are to be found on the Volume 4 microfiche (ED 092 605) under the title "Equating Tables, Error of Equating and Correlations, Grade 6".

Test Intercorrelations

Despite many reservations about the equating of reading tests with different content specifications the tests were all found to have high intercorrelations. Generally, the correlations for each test with each of the other tests fell little short of the correlation of that test with its alternate form. When the parallel-forms reliability estimates were used to obtain disattenuated correlations among the tests, very few of the correlations fell below .95, which is often used as an admittedly arbitrary cutoff for purposes of equating. Averaging across order of presentation, the disattenuated correlations for pairs of tests below .95 are listed in Table 3. All three cases at grade 4 involve the MAT, all four at grade 5 involve the SAT and all four at grade 6 involve the STEP. None of the disattenuated and averaged-over-order correlations among reading tests fell below .89 and the tests with low correlations changed from one grade level to the next. Although I agree with the judgment made by the investigators that the correlations are sufficiently high to justify the equating in all cases one is left with a curiosity about the tests that are involved in the "low" correlations at each grade.

In the case of the STEP test at grade 6 it may be that the "low" correlations are attributable to the difficulty level of STEP being somewhat out of phase with the other tests. Among the 7 tests in the original study for which the test intercorrelations are available, STEP is the only test that doesn't change levels during the 4th to 6th grade interval and by the spring of grade 6 STEP is an easy test relative to the other tests. Partial support for this interpretation can be found in Lord (1974). Despite the high intercorrelations of the tests Lord found the 7 tests in the original study to have fairly different patterns of relative efficiency at different percentile ranks. STEP was the only test to have higher relative efficiency than the MAT's at low percentile ranks but lower relative efficiency at middle and high percentile ranks.

Error of Equating

An important aspect of the equating design was the provision that made possible empirical estimation of the error of equating. This is accomplished by the use of McCarthy's balanced half-sample replication method (1966). The equating design consisted of a set of eight balanced half-samples. These half-sample replications were used to compute the root-mean squared deviation of the MAT equivalent scores for each half-sample replication about the MAT equivalent scores for the full sample. These errors of equating were computed for each of the eight combinations of methods and procedures and provided a means of judging the relative quality of the methods. The estimated error of equating also provided a basis for judging the overall adequacy of the equating for each test. For the preferred equating procedure and method (i.e., the average of procedures 1 and 2 and the equipercentile method) the estimated error for all tests was generally less than one

Table A-3

Pairs of Total Reading Tests with Disattenuated Correlations Averaged over Order of Presentation Below .95

(Value of correlation reported in parentheses)

Grade 4		Grade 5		Grade 6	
MAT-CAT	(.94)	SAT-STEP	(.89)	STEP-CAT	(.946)
MAT-ITBS	(.93)	SAT-CTBS	(.92)	STEP-CTBS	(.91)
MAT-SRA	(.93)	SAT-CAT	(.94)	STEP-ITBS	(.92)
		SAT-SRA	(.94)	STEP-SAT	(.93)

raw score point (substantially so in most cases). The only major exception to this is for test scores in the "chance" range. Based on these error of equating estimates, the equating would seem quite satisfactory for most practical purposes.

Comparison to Publishers' Norms

Once the tests were equated the norms obtained for the MAT were used to convert equivalent raw scores on all other tests to percentile ranks. Thus, the anchor test norms can be used to obtain nationally representative norms for all of the tests. With norms for all tests in hand, the next natural step was to compare the ATS norms to the norms provided by the publisher. The maximum difference between the ATS percentile rank (PR) of any test score and the PR of that same score on the publisher's norms is listed in Table 4 for each test at each grade. Also summarized in Table 4 is the typical sign of the ATS PR minus the publisher's PR for scores above and for scores below the median. A plus sign indicates that a given raw score would typically have a higher PR on the ATS norms than on the publisher's norms. In other words, a given score would appear better according to ATS norms than publisher's norms where there is a plus sign. The converse is true of a minus sign and a zero indicates that there is not a consistent difference in that the PR's are essentially equal.

As can be seen in Table 4, the maximum difference is relatively small for most tests at most grade levels. The SAT, and to a lesser extent the GMT (grades 4 & 5) and the MAT (grade 4) are notable exceptions to this statement. The differences for those tests are substantial. It may be of interest to note that the GMT and the SAT are the oldest of the eight tests. As indicated in Table 1 the SAT and GMT used in the ATS were both 1964 editions. It should also be noted that since the ATS was undertaken a new edition of the SAT has been published. (Harcourt Brace Jovanovich, 1973). Thus, the large differences for the SAT are somewhat irrelevant. The other large difference (MAT grade 4) may be attributable to the fact that separate answer sheets were used in the ATS whereas the publisher's norms at grade 4 are based on scorable test booklets.

For use with the interpretation of individual scores most differences between publisher's and ATS norms are not large enough to cause problems. If someone is interested in evaluating trends for groups of students, however, changing from publisher's norms to ATS norms might make quite a noticeable difference. To get a better fix on implications of changing to ATS norms for group data it would be desirable to have a table like Table 4 showing the differences between ATS school mean norms and publishers' school mean norms. Not all publishers provide such norms, however.

Subgroup Results

The tests were not only equated for the total sample but also for eleven special subgroups resulting from four breakdowns of the

Table A-4

Summary of Comparisons of ATS Norms

with Test Publishers Norms

Test	Grade	Maximum Difference in Percentile Rank			Typical Sign of ATS Minus Publisher's Rank	
		Vocabulary	Comprehension	Total	Below Median	Above Median
CAT	4	2	3	3	-	0
	5	3	2	3	-	0
	6	4	2	3	-	0
CTBS	4	4	2	3	-	0
	5	3	3	3	-	+
	6	4	6	5	+	+
GMT	4	3	10	*	0	+
	5	3	8	*	0	+
	6	3	4	*	0	0
ITBS	4	5	5	*	+	+
	5	6	7	*	+	+
	6	6	7	*	+	+
MAT	4	3	3	2	+	0
	5	3	2	3	+	0
	6	3	3	2	+	0
STEP	4	*	*	5	+	+
	5	*	*	5	+	+
	6	*	*	4	+	+
SRA	4	5	3	3	+	+
	5	5	2	3	+	+
	6	4	2	2	+	+
SAT	4	8	11	*	-	-
	5	15	12	*	+	+
	6	18	16	*	+	+

*Publisher's norms not provided.

sample on the basis of sex, SES, IQ, and race. For the sex breakdown no major differences were found. The results for the three IQ groups showed some differences but generally the differences were small except in regions where the data were relatively sparse. Thus, the total group equating tables appear satisfactory regardless of sex or IQ level.

The results of SES and for race were less similar. There was a consistent tendency at all grade levels for the high SES children to score higher on the CTBS than on any of the other tests and for low SES children to score lower on the SRA than any other test.

Marked differences in equating lines were also found for sub-groups formed on the basis of race. This is particularly true for the Spanish-surnamed sub-group which tended to score consistently lower in the top part of score range on the ITBS and SRA than on the other tests. The deviations for the black sub-group were not as large as for the Spanish-surnamed sub-group. Furthermore the deviations for the black sub-group were not consistent over all grades. There is some tendency at the upper score ranges, however, for blacks to score higher on the CTBS and SAT than on other tests at grade 4 and to score higher on the ITBS than on other tests at grades 5 and 6.

Although the sub-group equating results are undoubtedly the most provacative of the entire study it must be noted that "...the study was not explicitly designed to yield stable equating relationships for the minority sub-group children..." (ATS. Final Report. Project Report, p. 196). The sample size for the minority groups is extremely small in the parts of the score range where the largest differences were observed. Hence, the advice of the project report against using the racial sub-group equivalency score data is probably sound. But, this is an area of concern that deserves more intensive study and such work is currently under way (John Bianchini, personal communication).

UTILITY

The Transfer Student

In the announcement of the ATS contained in the fall 1974 issue of ETS Developments (ETS, 1974) a hypothetical girl named Mary is described. Mary and her parents moved. Her "new" school uses the ITBS but her old one used the STEP. Thanks to the ATS, Mary's new teacher can convert Mary's raw score on the STEP Reading to an equivalent raw score on the ITBS Reading. It might be added that either of these raw scores can be interpreted in terms of the national norms provided by the ATS.

Although the above claim is true it assumes that the teacher will (1) know about the ATS and (2) have the equivalency tables available. Both of these assumptions seem questionable to me. A major effort would be required to make this type of information broadly known by teachers. One way of accomplishing the goal might be for the publishers

to do the conversion to ATS percentile ranks for the users, and indicate the tests for which the percentile ranks are equivalent. Without such heavy use by publishers, however, I doubt that Mary's teacher would know how to convert Mary's score even assuming that she received raw scores rather than grade equivalents or some other standard score for Mary.

The need for publisher involvement to make the ATS results maximally useful prompted me to write to the six publishers that produce the eight tests involved in the ATS to ask about their plans. In the fairly limited time between my letters to publishers and the writing of this review I received responses from four of the six publishers. None of these four publishers plans to routinely provide ATS norms to their users. But, they all plan to make the information about the study available by informing their sales representatives and/or describing the study in their publications.

The limited effort on the part of publishers to make ATS norms and equating results known may be as much as could be expected of the publishers. It seems doubtful to me, however, that the planned level of effort will be sufficient to get a very large segment of the test users (including Mary's teacher) to use the ATS results.

By way of explanation of their limited plans to use the ATS results the publishers cited several practical limitations of the results. These limitations included: (1) the lack of data for tests other than reading, (2) the lack of data for grades other than 4, 5 and 6, (3) the lack of data for the publisher's alternate forms, and (4) the lack of scaled scores. All of these factors were viewed as limiting the practical value of the ATS results for their users.

Changing Tests

Schools are sometimes slow to switch from one test to another because of experience with one test and the comparative value of the historical data. The ATS results make it possible to make a change and still have the ability to compare current reading test results to historical results in terms of the ATS norms. Again this assumes that the knowledge of this capability is available to the school.

Measuring Change

Another use that has been usggested for the ATS data is in the measurement of change where one publisher's test is used at time 1 and another publisher's test at time 2. Presumably this could be done in terms of percentile ranks. This might be appropriate for gauging the direction of change in relative standing as suggested by Coleman and Karweit (1970) but not for estimating the magnitude of change. There are major differences between change as measured in terms of percentile ranks and as measured in terms of a vertically equated scale such as grade equivalents (see for example Linn, 1974).

The ATS was not designed to vertically equate tests that change levels from one grade to the next. It does provide some indirect information for this purpose, however. For example, the same level of the CAT was used at grades 4 and 5 but different levels of the MAT were used at those grades (see Table 1). By using the CAT equivalencies of the MAT it is possible to convert the MAT Elementary Level Reading scores to equivalent Intermediate Level Reading scores. There are a number of other tests with a constant level over grades 4 and 5 that might be used for this purpose and for the best estimate it would be desirable to use some sort of combination of the various estimates. For purposes of illustration, however, I selected a few scores of the CAT at grade 4 and noted the equivalent Elementary Level MAT scores. The same CAT scores were then used at grade 5 to find the equivalent Intermediate Level MAT raw scores. These scores are shown in Table 5. Finally, the publisher's norms were used to convert the equated MAT Elementary and Intermediate raw scores to grade equivalent scores. The resulting grade equivalent scores are also reported in Table 5.

If the two columns of grade equivalent scores in Table 5 are compared some non-trivial differences in the grade equivalents can be observed. The largest of the differences in corresponding grade equivalents shown in Table 5 occurs for MAT raw scores that are equivalent to a CAT raw score of 60. At this level the grade equivalent scores are 6.6 at grade 4 and 7.4 at grade 5 for a difference of .8 grade equivalent units which would presumably be interpreted as almost a "year's gain." Throughout the range the grade equivalents tend to be larger at grade 5 than at grade 4.

The above analysis in terms of grade equivalent scores is admittedly rather crude and does not begin to scratch the surface of the number of possible comparisons of this type that might be made. It is not intended to imply that growth should be measured in terms of grade equivalent units, in fact, I have elsewhere argued to the contrary (Linn, 1974). Furthermore, the results in Table 5 may be an artifact of the nature of grade equivalent scores and they are not the score unit to use in equating. But, the person who is interested in measuring change needs some sort of common score and will usually want something besides percentile ranks. If so, some form of the publisher's scaled scores is still the natural recourse. The above analysis suggests that the results of such comparisons may be very misleading at least if grade equivalent scores are used.

Aggregation of Results from Several Tests

Possibly the most significant use of the ATS may come from making it possible for a governmental agency to aggregate reading test scores across several tests. This is a potentially important use in that it conceivably could greatly reduce the need for special test administrations for information purposes at the state or national level. As noted previously programs such as Title I ran into considerable difficulty in

Table A-5

Total Reading Equivalent Scores on the MAT

Elementary and Intermediate Levels

Level 3 CAT Raw Scores (Grades 4 & 5)	Equivalent MAT Raw Scores and Corresponding Grade Equivalents			
	Elementary Level (Gr. 4)		Intermediate Level (Gr. 5)	
	Raw Score	Grade-Equivalent	Raw Score	Grade Equivalent
80	94	9.9	91	9.8
70	89	8.4	76	8.4
60	84	6.6	63	7.4
50	76	5.2	51	5.5
40	63	3.7	39	4.4
30	45	3.2	29	3.5
20	26	2.3	20	2.6
10	12	1.3	8	1.4

trying to make sense out of test score data from a wide variety of tests. State agencies have had similar problems which has led to the use of single tests for statewide testing in some cases. Thanks to the ATS results schools should be free to select their own reading test from among the eight involved in the ATS while the capability of aggregating data at the district, state or national level is still maintained.

I would not find it surprising if aggregation is the main use that is made of the ATS results. After all, it was the desire to have this capability that made the ATS a reality after over 30 years since Cureton (1941) made his plea for an anchor test study.

LIMITATIONS

In my opinion, the ATS is an extraordinarily sound study from a technical point of view. Most of the limitations, some of which have been implicitly noted above, come about more from the scope of the study than from the implementation. There are three rather obvious limitations of this nature that I would like to mention at this stage. These are (1) test content, (2) grade levels, and (3) the absence of vertically equated scaled scores.

Although reading would probably be most people's first choice if a single content area is to be involved, there are obviously other important content areas. Many would argue that even a complete achievement test battery puts the focus on much too narrow a range of educational goals. By making it possible to aggregate only for reading tests the emphasis becomes even narrower. Although equating of tests in other content areas may be desirable it would be unreasonable to expect one study to do everything and the ATS is already a giant. Furthermore, the technical feasibility of equating in other areas may be limited due to less similarity in what is measured in content areas other than reading, from one test battery to the next.

The choice of grades 4, 5 and 6 was partially based on high test usage at those grades. They are a reasonable starting place but the same problems that prompted the ATS remain unresolved at other grade levels.

The absence of an effort to vertically equate tests that change levels in grades 4, 5 and 6 and create a common scaled score is regrettable from my perspective. Without doing this the test user who wants to analyze scores across levels must revert to the publisher's norms. As good as the publisher's norms may be, they do not live up to the ATS standards.

I also think that the absence of a common scaled score is a missed golden opportunity. By creating a new scaled score that is common to all tests it might have been possible to reduce the diversity in types of scaled scores which confuse users and more importantly to speed the

demise of some undesirable types of scores. In this way the ATS might have helped achieve standard D5.2.3 of the 1974 Standards for Educational and Psychological Tests (APA, 1974). According to standard D5.2.3 "Interpretative scores that lend themselves to gross misinterpretations, such as mental age or grade-equivalent scores, should be abandoned or their use discouraged. Very Desirable" (APA, 1974, p. 23). The absence of scaled scores could be rectified through secondary analysis of the data. The data that are required are available.

A final limitation that I'd like to mention has to do with time rather than scope. As noted above, one of the test batteries (The SAT) has already been revised. This is apt to happen to several of the others within the next 5 or 6 years. In view of this it seems unfortunate that there was a delay of almost two years between the completion of the final report and its release by USOE.

CONCLUDING REMARKS

The ATS is a landmark study. It is a tribute to careful planning, superb execution and high technical capability. The goals of obtaining representative norms and equating several widely used reading tests at grades 4, 5 and 6 were clearly accomplished. So too, were the several minor goals. The results of the study should prove to be of considerable practical value especially to governmental agencies that want to aggregate scores across several tests. The data bank which was created by the study should be valuable for a number of secondary analyses.

Despite these major accomplishments, one need only look back at Cureton's original plea for an anchor test study to realize that there is a long way to go to achieve his ideal. According to Cureton, "An ideal system of norms should be based on a specially constructed and standardized test, and its units should be stable from year to year, from test to test, and from early childhood to old age. They should also be as directly meaningful as possible in terms of the existing concepts of the population in general and the teaching population in particular...The ideal anchor test should yield separate scores for all the major intellectual factors in the school achievement complex" (1941, pp. 291-292). We clearly have a ways to go. Given the expense of equating tests of reading at three grade levels and the fact that other content areas and other grade levels pose more difficulties it seems doubtful to me that we will achieve Cureton's goal.

REFERENCES

American Psychological Association, Standards for Educational and Psychological Tests, Washington, D. C.; 1974.

Angoff, W. H. Equating non-parallel tests, Journal of Educational Measurement, 1964, 1, 11-14.

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.) Educational Measurement 2nd Edition, Washington, D. C.: American Council on Education, 1971.

Coleman, J. S. & Karweit, N. L. Measures of School Performance, Santa Monica, California: Rand, R-488-RC, July 1970.

Cureton, E. E. Minimum requirements in establishing and reporting norms on educational tests. Harvard Educational Review, 1941, 11, 287-300.

Educational Testing Service. ETS Developments, Princeton, New Jersey, Educational Testing Service, 1974, 21, No. 4.

Flanagan, J. Equating non-parallel tests. Journal of Educational Measurement, 1964, 1, 1-4.

Harcourt Brace Jovanovich, Stanford Achievement Test, 1973 edition. New York: Harcourt Brace Jovanovhic, 1973.

Jaeger, R. M. The national test equating study in reading (The anchor test study). Measurement in Education, 1973, 4, 1-8.

Lennon, R. T. Equating non-parallel tests. Journal of Educational Measurement, 1964, 1, 15-18, (a).

Lennon, R. T. Norms. In Proceedings of 1963 Invitational Conference of Testing Problems, Princeton, New Jersey: Educational Testing Service, 1974 (b).

Lindquist, E. F. Equating non-parallel tests, Journal of Educational Measurement, 1964, 1, 5-10.

Linn, R. L. The use of standardized test scales to measure growth. Conference of Policy Research: Methods and Implications, University of Wisconsin, Madison, Wisconsin, May 1974.

Lord, F. M. Quick estimates of the relative efficiency of two tests as a function of ability level. Journal of Educational Measurement, 1974, 11, 247-254.

Loret, P. G. The anchor test study. Paper presented at the 1974 Arizona Education Association -- Education Fair, Phoenix Arizona, Oct. 31-Nov. 1, 1974.

McCarthy, P. J. Replication: An Approach to the Analysis of Data from Complex Surveys, Washington, D. C.: National Center for Health Statistics, Vital and Health Statistics, Series 2, No. 14, 1966.